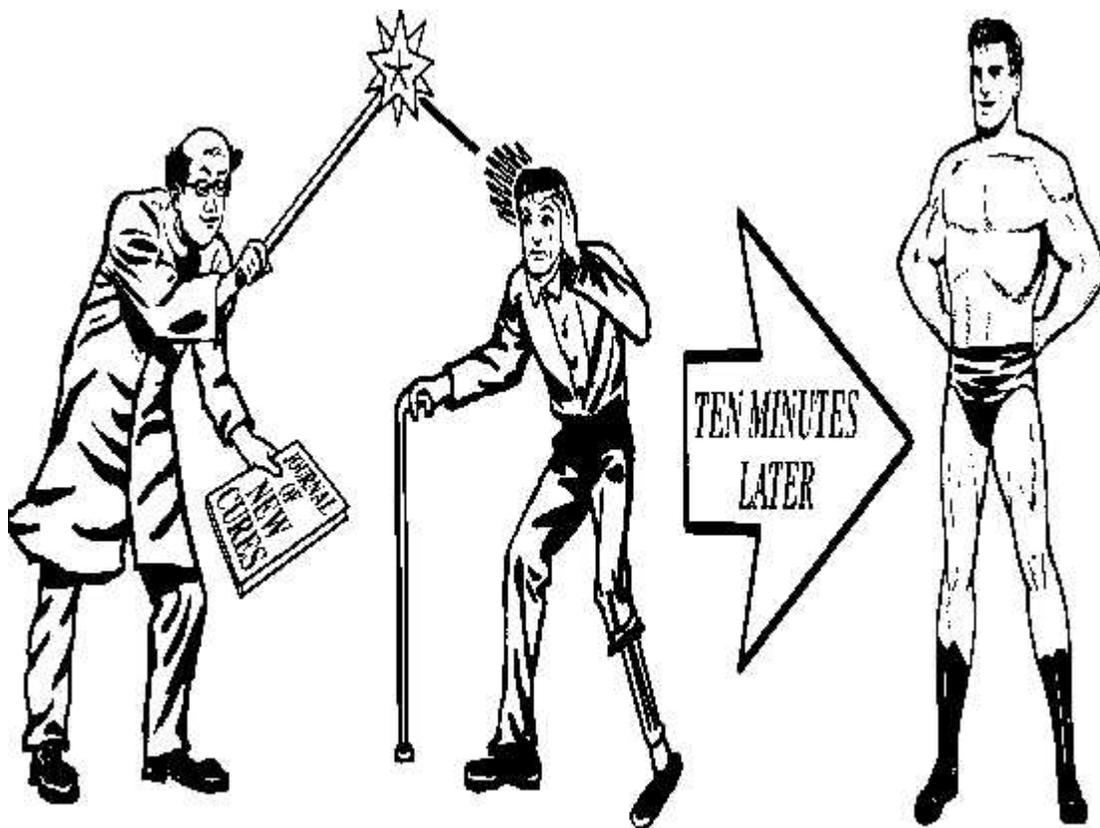


Clinical Research:

Skills clinicians need to maintain effective practices

By Richard A. Sherman, PhD



The art and science of establishing clinical credibility.

Or, how can you tell if that amazing new treatment really works?

Cardinal Rules for Establishing Credibility When You:

Prepare a <i>Clinical</i> Presentation / Article	Listen to / Read a <i>Clinical</i> Presentation / Article
<p>Be certain to:</p> <ol style="list-style-type: none"> 1. Title your presentation / article appropriately so it doesn't promise more than it can deliver or ascribe changes to one aspect of a multifaceted intervention. 2. Begin with a brief summary of what you did / found. 3. Describe the general characteristics of the group you worked with and define your inclusion and exclusion criteria. 4. Present how your patients were diagnosed. Don't fall into the trap of trusting diagnoses by others if such diagnoses are known to be frequently incorrect (e.g. physicians are terrible about correctly diagnosing headaches). Use recognized criteria so your audience will believe that your patients had the disorder you claim to be treating. 5. Use correct assessment techniques for the disorder (e.g. the MMPI is not valid for establishing the psychological components of low back pain). 6. Define your assessment so you establish the basis for saying that people learned the tasks you were teaching during your treatment. E.g., if you are teaching people to change their muscle tension, show the baseline status then show that those people who improved changed in the desired direction. This establishes the relationship between the intervention and changes in symptoms. 7. Use the correct outcome measurers and use them correctly. Review the literature so you are up to date. For example, <i>0 – 10 analog pain scales must define 10 to have an objective limit such as “would faint if had to bear the pain for one more second”</i> rather than “most pain can imagine”. 8. Establish pre and post treatment baselines of sufficient duration to establish symptom variability. E.g. headache baselines need to be between two weeks and a month. This is how you demonstrate effectiveness. 9. Use the correct design for the level of work done on the intervention already. E.g. a new idea needs only a baseline – intervention – baseline design while a test of an idea which has been shown to produce changes needs to incorporate a control group to show that changes are not due to non-specific effects. 10. Include sufficient subjects so your results are likely to be due to the intervention rather than chance variability. 11. Clearly explain what your intervention was and have some way to know that there was sufficient intensity to have a chance of causing a change. E.g. one relaxation session isn't likely to cure anything. 12. Present your results clearly with graphics rather than just tables. Show sufficient descriptive statistics so people decide what happened. 13. Never ascribe symptom changes to one aspect of a multifaceted intervention when you have no way to tease out the effect of that aspect. E.g. if you gave relaxation training and biofeedback, don't say that the changes were due to biofeedback. 14. Not worry about the need to prove an underlying mechanism for the technique you used. All you need to do in a clinical presentation is demonstrate that a change did take place. Other types of research demonstrate how & why. 	<p>Is / are there:</p> <ol style="list-style-type: none"> 1. Adequate diagnosis and assessment of the subjects? 2. Adequate pre treatment baseline to establish symptom variability? 3. Objective outcome measures relevant to the disorder? Were they used correctly? 4. Intensity of the intervention sufficient to produce an effect? 5. Way to check whether the intervention was successful (drug taken properly, behavioral technique successfully <u>learned</u> and then used). 6. Sufficient patient-subjects so result is credible? 7. Appropriate design for the question (e.g. single group, controls, believable placebo, etc.?) 8. Sufficient descriptive statistics so results are clear? 9. Long enough follow-up so duration of results can be established? 10. In a multifaceted intervention, were any changes in symptoms ascribed to one element of the intervention when there is no way to differentiate the effects of each part?

Contents

Inside Front Cover	
Table on cardinal rules for establishing credibility	2
Introduction	6
1. Rationale - the crucial need to understand and use clinical research	6
2. The scientific method	12
Section A. The need to know what you are doing	14
Chapter 1. The basic steps and time-line of a project	14
Chapter 2. Defensive reading of clinical literature - does the hypothesis make sense?	22
Chapter 3. Protocol development - formulating and maturing a question	25
Chapter 4. Background and literature searches	37
Chapter 5. Determining feasibility	40
Chapter 6. Research ethics	42
Chapter 7. The research protocol approval process	69
Chapter 8. Pitfalls in the first steps – can you trust interpretation of the data by people who are not neutral?	74
Practical exercises for Section A	76
Section B. Basic study structures for the office and clinic environment	78
Chapter 9. The logic and progression of designs	79
Chapter 10. Exploratory single subject and single group designs	87
Chapter 11. Observational studies - longitudinal & cross-sectional designs	91
Chapter 12. Prospective experimental study designs	94
Chapter 13. Outcome and quality of life studies	100
Chapter 14. The protocol's research plan and design	104
Chapter 15. Defensive reading of clinical literature - does the design fit the needs?	106
Chapter 16. Pitfalls in study design	108
Practical exercises for section B	112
Section C. Establishing the credibility of data and clinical publications	113
Chapter 17. Subject selection techniques - sampling, inclusion - exclusion	114
Chapter 18. Hardening subjective data	119
Chapter 19. Validity and reliability - defensive data entry	128
Chapter 20. Survey, test, and questionnaire design	137
Chapter 21. Defensive reading of clinical literature - can you trust the subjects and data?	152
Chapter 22. Pitfalls in data gathering methodology	153
Practical exercises for section C	156

Section D. Statistics for evaluating literature & interpreting clinical data	157
Chapter 23. Concepts of clinical data analysis	157
Chapter 24. Descriptive statistics for evaluating clinical data	161
Chapter 25. Probability and significance testing	171
Chapter 26. Decision / Risk analysis and evaluating relative risk	174
Chapter 27. Power analysis - determining the optimal number of subjects	180
Chapter 28. Evaluation of overlap between groups through inferential statistics	187
Chapter 29. Evaluation of relationships between changing variables	195
Chapter 30. Dichotomous and proportional data	205
Chapter 31. Outliers - data points that don't meet expectations	210
Chapter 32. Pattern analysis	212
Chapter 33. Survival / life table analysis	215
Chapter 34. Establishing Efficacy	218
Chapter 35. Establishing Cause and Effect	225
Chapter 36. Defensive reading of clinical literature: Handling the data	242
Chapter 37. Pitfalls in study analysis	247
Practical exercises for section D	250
Section E. Synthesizing the elements to produce an effective study	252
Chapter 38. The protocol - incorporating statistics	252
Chapter 39. The grant - extramural funding process	255
Chapter 40. Writing and presenting the study	260
Chapter 41. The publication submission and review process	267
Chapter 42. Changing clinical practice based on what you did and read	273
Chapter 43. Defensive reading of clinical literature - does the conclusion match the raw data, the data analysis, the hypothesis, and the background?	275
Chapter 44. Pitfalls in the overall research process	279
Practical exercises for section E	280
Section F. Glossary of research and statistical terms	281
Section G: Typical protocol format and structure	300
Section H: Sample protocols and papers	308
Sample 1. Protocol and consent form for a pilot study	308
Sample 2: Protocol and consent form for a controlled study	326
Sample 3. Unacceptable protocol and consent form	337
Sample 4. Protocol using non-human animal subjects	342
Sample 5. Poor paper	354
Sample 6. Paper based on pilot study in sample 1	359
Sample 7. Grant application based on pilot study in sample 1	366
Section I: An algorithm for the steps in evaluating clinical protocols and articles	384

Section J: Can you trust the interpretation of study results by people who have not adequately tested their assumptions or are not neutral?	388
Section K: Use of Effect Size Calculations to determine efficacy of biofeedback based interventions	392
Section L: Further reading and References	448
Publishing and copying information, about the author	453

Introduction

1. Rationale:

A. So, you're a "dyed-in-wool," "carved-in-stone" clinician who knows there is always room for improving your skills and that innovations in practice are always thundering urgently over the horizon. The question is how to figure out which of those shiny new "innovations" you are deluged with warrant changing your practice or taking a course to learn some new skills in order to apply them. In other words, how do you separate the wheat from the chaff - or the good ideas which are still half baked from those ready to eat? How do you know your own interventions are effective - or if they could be even more effective? That's what this book is all about.

The bottom line is that you need to have a working grasp of the knowledge, skills, and concepts associated with research in order to:

1. Develop, extend, and change your clinical practice through knowledgeable assessment of the clinical literature and presentations. In other words, You need to be able to decide whether to change your clinical practice based on what you read and hear.
2. Track your ongoing clinical work to spot the holes in daily practice and determine which interventions are effective on both a "case-by-case" and "disorder-by-disorder" basis.
3. Push the envelope of your specialty forward and expand the use of your clinical skills to untried disorders without exceeding your resources in such a way that both you and your colleagues are convinced that your approach is actually effective.

You need to be able to incorporate the well demonstrated techniques of research into your decision making process because real clinical life is unfocused, foggy, murky, and uncertain. Research is an attempt to weave tendrils of mist and fog into a graspable fabric. Research does this by optimizing the ability to focus and clarify the smidgens that can be distinguished.

The most emotionally evocative "picture" of the pathos and angst of life in the clinical environment I have seen was painted by E. Nelson Clark when he said, "The treatment of flexion contractures in the digits is a varied and often Machiavellian science. Individuals attempting to rehabilitate the hand have an arsenal of weapons cunningly and scientifically designed to straighten the dreaded painful, swollen, and bent finger. Sadly, and all too often, our clients patiently suffer through all of the gadgetry, bells, and whistles, only to be left with a swollen, painful, and still bent finger. Unfortunately, the selection of splints may often be inappropriately based on tradition, fads, bad science, or a tribute to their creators." It is the intention of this book

to provide you with the tools you need to wend your way through this all too typical morass toward optimal decisions.

B. Every clinician should consider actually performing at least one research project even though they just want to be clinicians and never plan to get near anything that even looks like research for the rest of their lives because:

1. Virtually no research projects actually work the way they are supposed to. Nobody would believe the difficulties and oddities that arise when actually doing what appears to be even the simplest project. Until you try it yourself, you will never truly believe it. You must have this belief ingrained into you before you try to assess the clinical literature because this is how you grow alarm bells for a study that is just too good to be true.

2. Most people tend to paint life with as broad a brush as possible. Actually performing a research project forces you to look at the details of the process. All of the lumps and bumps that can't be seen when watching a nature show on TV become very obvious when you are on the spot.

3. Taking ownership of your specialty: When you do a research project, you advance knowledge and understanding of your specialty - even if in a minor way. That makes you a part of the field. It is now partially yours because you helped make it what it is.

What do you need to learn from the research process?

**You simply need to recognize when you don't know something
you need to know to get the job done!**

You need to know enough to know -

There is a gap in your knowledge when you come to one

Which holes in your knowledge base you can fill in enough to do the job

Which gaps you can bridge alone and when you need help from another specialty

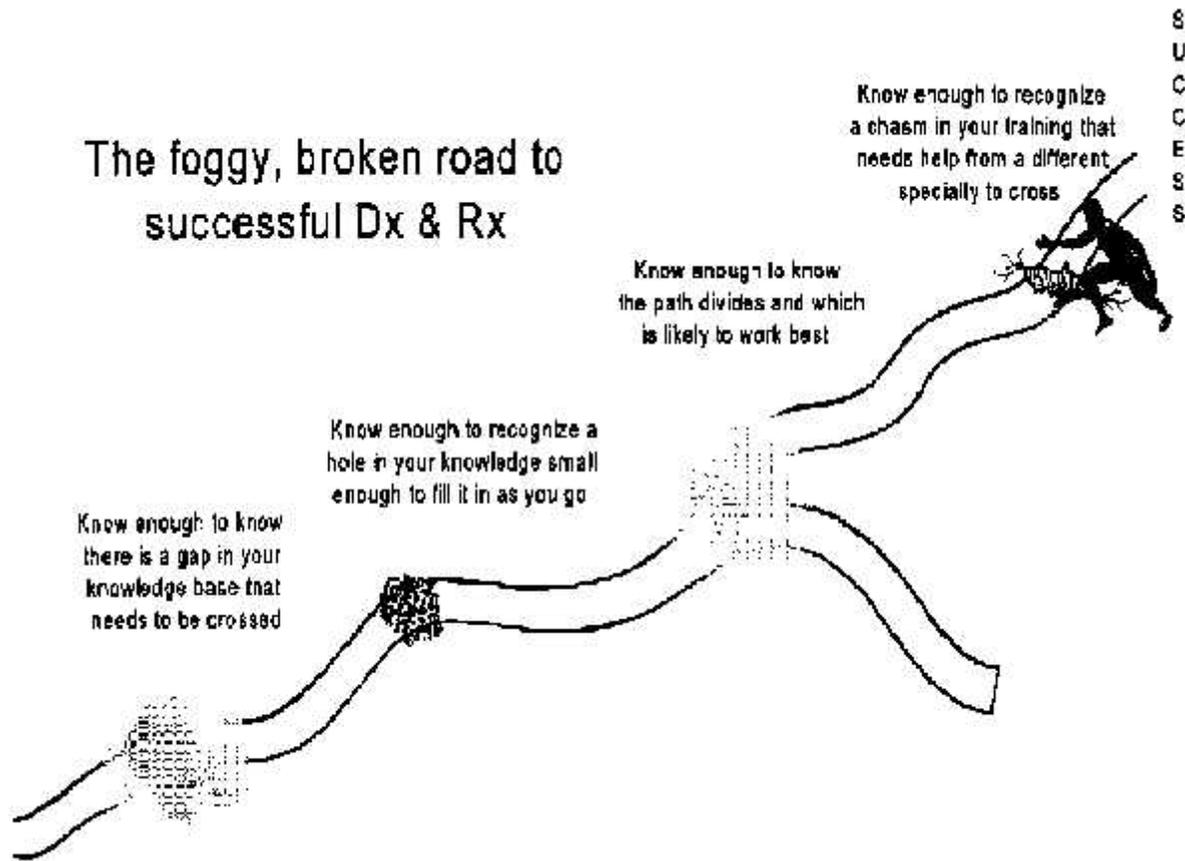
Which chasms have to be avoided

To differentiate a pothole from a gap from a chasm

This concept is illustrated in Figure 1.

If you were to incorporate the information from this book into your clinical work, what would happen? Your way of approaching your work would incorporate ideas which give you a logical basis for changing your clinical practice and deciding whether your therapies are effective.

Figure 1



D. Let's spend a minute considering how clinical practice frequently gets changed in relation to models of credibility. The typical models include:

1. The island fortress - in which a self reinforcing group tells each other how great their therapies are even if nobody else likes them. Questioners are thrust away and new believers are recruited with zeal.
2. The great person cult with priests and acolytes - in which a charismatic leader makes pronouncements and the flock accepts them not based on what was said but that the leader said it. Uncritical press coverage tends to be intense once the press becomes familiar with the leader.
3. The hard sell - in which drug and equipment providers push tentative findings far out of proportion and make sure everyone who could conceivably use or influence the use of the innovation hears about it in the most positive light. They frequently sound so convincing that clinicians alter their practices on the basis of incredibly flimsy evidence.

4. The word is out - in which the newest trends spread by word of mouth (and the Internet) and are met with uncritical enthusiasm. Word of unpublished trials are mistaken for proof and enthusiasts take off. In this model, the "establishment's" response to the innovation is to initially ignore or ridicule it as ludicrous, then to violently oppose it, and finally, if it becomes popular, accept it as efficacious because practitioners are using it.

In all of these models, the innovation is frequently never adequately tested and people waste decades doing no good what-so-ever for their patients. In those cases where no harm is done, abiding by Voltaire's dictum that "the art of medicine consists of amusing the patient while nature heals the disease" is fine. However, all too often real harm is done because patients delay getting adequate care until the disease is incurable or the treatment does real harm. Frontal lobotomies for housewives' depression and open heart massage for angina are just two of hundreds of well substantiated examples of untested therapies spreading like wildfire and only being damped out when objective research gradually shows them to be useless.

E. At about this point you may be saying to yourself: "That's ancient history - we've learned our lesson since then. Can there actually be whole clinical areas based on trash science and unsubstantiated, poorly done clinical findings which continue to plague us?" You know the answer - of course there are. One particularly pathetic, current example is the use of thermography to objectively detect and diagnose low back pain. The idea that patterns of heat on the surface of the back and legs might reflect physical problems underneath is not as irrational as it might seem. One would expect a change in blood flow (and thus, heat) to dermatomes in the legs concurrent with impingement of nerves serving those dermatomes. One would also expect inflamed low back muscles to give off more heat. Dozens of "clinical case series" style articles appeared which claimed to show discrete patterns of heat related to specific categories. Thermography burst into the courtroom as a way to objectively tell that the complainant actually had pain. Thermographers showed many hundreds of credulous juries color photos of asymmetrical heat patterns taken from the wretched patients' backs - and they usually won their cases. Various types of thermographs - some costing over \$70,000 began selling like hot-cakes. Courses in how to perform thermography mushroomed. A journal of clinical thermology even started. After nearly a decade, during which numerous clinicians claimed that they couldn't find the patterns, several grants were given to find out what was happening. I happened to get one from the VA. The project took us over two years to complete and showed that the patterns were random nonsense - mostly within the range of normal side-to-side variation. The findings from several other studies appeared at about the same time as ours. All of a sudden, the plethora of clinical studies dried up - and the field died with them in a few years. Unfortunately, this is really how ineffective clinical innovations are tested. First they spread and then they are stamped out - if they can be. Some gain enough believers on the fringe and are in sufficiently unregulated areas that they can continue forever if the money from credulous patients keeps coming in. Another example is the widespread use of bone marrow transplants used to prevent recurrence of breast cancer. The technique spread like wildfire for a decade without being tested and was finally shown to be a flop once the testing took place (Brownlee 2002). The bottom line is to know when to test a clinical innovation instead of dragging it along forever.

The other side to this issue is exemplified by Admiral Hyman Rickover's statement that

"Good ideas are not adopted automatically. They must be driven into practice with courageous impatience." In other words, you may have a good, valid idea but it takes work to get it accepted beyond your tiny group. Just now, what the Western World wants is objective evidence that a clinical technique works before it can join the political agenda that may lead to acceptance. You need to know how to get a reasonable level of evidence with a reasonable level of labor and resources if you are going to get a new technique to the point when it could be considered for acceptance. The political part of the process is considered in the ethics chapter.

F. Clinicians are gradually moving toward combining knowledge of research techniques with clinical skills to make practice decisions. For example, Carmine Iacono pointed out that use of evidence based medicine (EBM) is a growing trend in primary care medicine. It argues that we can no longer rely on clinical intuition alone (see Greenhalgh, 1997). EBM means integrating clinical expertise with the best available external clinical evidence from systematic research. Systematic research means controlled studies. This comes down to combining clinical judgement with relevant clinical research. EBM's basic principles include:

1. Develop good answerable clinical questions.
2. Track down the best evidence to answer the questions.
3. Critically appraise the evidence for validity and usefulness.
4. Apply the results to the clinical problem at hand
5. Evaluate the results

More information can be found in Greenhalgh (1997).

G. One last word about doing research: Actually performing research happens to give you something very important. It gives you permission to stop rushing about accomplishing tasks. To do a really superb job at research, every once in a while you need to stop and put your feet up, stare at the sky, take a walk - all guilt free. You get to do all these great things without feeling guilty because you are really busy as a bee letting your mind drift over your research concepts, letting it poke into dusty corners you didn't even know you had, letting it run over and over all sorts of maybe related stuff. You can't tell yourself you're not accomplishing anything. Why, you're doing the most important part of research of all - planning! So, kick back and join the research community.

So, how are you supposed to learn to do all this magic and make all these decisions?

As you might have guessed, just reading this book isn't going to do it. You need to think about and practice the concepts presented. Some of the concepts - especially in the ethics section are very controversial so need to be thought through very carefully. Practical exercises are presented at the end of each major section not only to give you practice handling the section's concepts but to potentiate your ability to incorporate the concepts into your professional life.

WARNING

Limited Depth Ahead

This book is intended for clinicians doing mainly clinical work with some simple basic studies to support that clinical endeavor. Thus, the material presented here matches the depth of knowledge about experimental design and statistical analysis required to meet these needs. If you need more information, you are probably not a typical clinician doing typical clinical studies or your are in deeper than you should go without professional support.

Another way this book attempts to help you transform the plethora of abstract concepts into practical terms is by presenting a typical study and following it the progression of its development. This is a real study which was attempting to determine whether a new treatment for migraine headaches actually worked. It come complete with all the flaws and problems which can be expected of any study. The author's names have been changed to protect the innocent. Material related to the sample study is presented in text boxes such as the following scattered throughout the text:

Sample Study

Prevention of migraine headaches by exposure to pulsing electromagnetic fields (PEMF)

The sample actually consists of two studies. A small pilot was conducted before the full fledged controlled study was attempted. Their protocols comprise samples one and two near the end of the book. You will be referring to these protocols as you read the first few sections of the book. When you get to the data entry sections, you will be creating a data set of the type that would have been gathered for the full study. You will then analyze the data set you create while plodding through the statistics section. Then comes the hard part - figuring out what the data mean with the help of the rest of the book.

If you don't have a data base program and a statistics program, you need to begin thinking about getting them so you are ready to begin working with the samples. If you are not familiar with these programs, please take a moment to read the "microwave" simile at the beginning of the statistics section (D).

Don't run out and spend a fortune on programs. All you need to follow this book is a data base of the type that comes with most computer packages such as lotus or excel and a very simple statistics program.

2. The Scientific Method:

The philosophy for performing and evaluating clinical research espoused by this book is entirely based on the "scientific method" so it should be discussed in some detail.

Just what is it? A very straightforward process for following a process as objectively as possible. Its basic steps are: Make an initial observation, develop a hypothesis (guess) to explain what was observed, test the hypothesis, and observe the results of the test. Note that this doesn't include asking "why" something happens beyond establishing a chain of events. The philosophical aspect of why things are as they are is not part of the scientific method but is certainly a part of science itself.

According to the Encarta Encyclopedia 2002 on line, "The era of modern science is generally considered to have begun with the Renaissance, but the rudiments of the scientific approach to knowledge can be observed throughout human history." Several historians credit early Arab scientists with developing the scientific method in about 1,000 ACE while others credit workers during the European renaissance somewhat later, but it has actually been around at least since the ancient Greeks. It is very likely that it was used in Ur in 4,000 BCE or even earlier.

The key component of the scientific method is to develop and use methods to objectively record observations without permitting preconceived views to influence the results or methodology. It is crucial that the methodology be so explicitly laid out that the observations can be repeated by others. In fact, within the scientific community, it is implicit that a set of observations will not be accepted until they are independently repeated.

Objective observation is the most important element of the method but not the only one. Once the initial observations are made, inductive reasoning (forming generalized hypotheses / theories from observations) must be applied to try to put the observations within the context of other knowledge and to design methods to test whether the guess about the relationship is correct. These "tests" are essentially experimental designs which attempt to predict how the observed data will change if the situation is modified. Deductive reasoning (reasoning from theories to account for the results of an experiment) is applied to the results of the intervention to explain what actually happened as opposed to what was originally predicted.

How does science relate to superstition? For purposes of this book, we can consider "science" to be the application of the "scientific method" to the evaluation any idea. In this context, it would be objective evaluation of clinical techniques. Superstition is the acceptance of unverified phenomena as real - in this context, accepting clinical ideas without testing them. Thus, it is hoped that science, with the "scientific method" as its bastion, stands firm against superstition and can be used to help us differentiate the wheat from the chaff in our confusing world of clinical practice.

Dr. Anthony Carpi (Quoted From the City U of NY's Web Site, 2003) has written a very fine explanation of the elements of the scientific method which I am repeating here. "A common misperception of science is that science defines "truth". Science does not define truth, but rather

it defines a way of thought. It is a process in which experiments are used to answer questions. This process is called the scientific method and involves several steps:

1. **Observation:** The first step of the scientific method takes place when an observation is made regarding some event or characteristic of the world. This observation might lead to a question regarding the event or characteristic.
For example, you might drop a glass of water one day and observe it crashing to the floor near your feet. This observation might lead you to ask a question, "Why did the glass fall?"
2. **Hypothesis:** In attempting to answer the question, a scientist will form a hypothesis (or some would say a guess) regarding the question's answer.
In our example there are many possible hypotheses, but one hypothesis might be that an invisible force (gravity) pulled the glass to the floor.
3. **Experimentation:** Of all the steps in the scientific method, the one that truly separates science from other disciplines is the process of experimentation. In order to prove, or disprove, a hypothesis, a scientist will design an experiment to test the hypothesis."

This book is devoted to showing you how to apply the scientific method to behaviors in order to strengthen your ability to make sense of them.

Section A.

The need to know what you are doing

Chapter 1

The basic steps and time-line of a project

A. Clinical Innovation - Where do clinical research projects come from?

Most clinical research starts with a question about a technique's efficacy. Frequently there is no information about the long term effectiveness of a commonly used technique or how well a diagnostic procedure actually reflects the magnitude of the problem. For example, it may be well established that a particular artificial hip remains solidly in place for years with few complications or failures but nobody may have asked the recipients how effective their artificial hips are in helping them walk. Many times, a clinician has thought of a modification of a current procedure, a novel application of a standard procedure, or has had a virtually new idea. If it doesn't appear that the innovation will do any harm or interfere with or delay standard treatment, most clinicians will go ahead and try the idea with a few carefully selected patients. If the idea seems to work, the clinician may try it on a few more less carefully selected patients and try to follow them longer than is usual in clinical practice. For example, when people break their collar bones, the broken ends are usually left to heal in whatever position they happen to be in. This frequently results in a large, sharp lump under the skin which interferes with wearing harnesses for packs, parachutes, seat belts, etc. An orthopedic resident (Greer Noonburg of Madigan Army Medical Center in Tacoma Washington) realized that a standard device for holding broken leg bones together (an external fixator) could be applied to the clavicle with out much chance of complications. Standard treatment would not be delayed because standard treatment just consists of putting the arm in a sling. He tried the idea with several patients whose clavicles' would have healed with sharp bumps and whose jobs required that they wear harnesses. They healed rapidly, without complications, and without lumps. He tried several more patients with the same result and followed them for several months to insure that no short-term problems developed. At this stage he decided that it was time for a formal pilot study so enough data could be gathered to

determine whether a full study was warranted and to begin establishing the actual efficacy of the technique. He presented his initial results to wide acclaim and it looks like the technique will become accepted, reimbursable, standard practice within a year.

The point to this is that clinical innovation is normally the starting point in the process of changing clinical practice. People tend to get very upset when these innovations are incorporated into general practice without objective testing and mix this error with the practice of clinical innovation itself. So, don't be a party to throwing the baby out with the bath water. Clinical innovation is the crucial first step in starting the process of changing clinical practice

Problems ensue when clinical innovations spread directly to clinical practice without the crucial pilot, and possibly formal studies, required to establish their efficacy.

Sample Study

The Crucial clinical innovation:

If you haven't read the first two sample protocols, please read them before proceeding further.

A protocol on exposing the inner thighs to pulsing electromagnetic fields to prevent migraines??? You have to admit that this is a really improbable topic. Why would anyone even think of trying to treat migraine headaches by exposing someone's thighs to pulsing magnetic fields in the first place??????

Well, this is what really happened. Several patients among a group receiving pulsing electromagnetic field (PEMF) therapy for femoral stress fractures reported that their migraine headaches has ceased during therapy. Coincidence??? Who could tell? It was already demonstrated that exposing thighs to PEMFs has no known negative side effects so the clinicians asked several of their other patients who happened to have well diagnosed, uncomplicated migraines known to occur at least four times per month if they would mind having their thighs exposed to PEMFs while their regular treatments for their orthopedic problems were in progress. The headaches went away and stayed away for at least several months!

On the basis of this finding, it was decided that a formal pilot study was warranted. They did not try their innovation on people not already under their care nor did they begin trying the innovation in their general practices.

B. Getting started

1. If you decide you are going to do research, what do you do to get started? In other words, **What should you ask?** Or - where do research ideas come from? What's a good idea to

do research on?

Key take home point: The closer you remain to your basic clinical interests, the more likely it is that you will allocate the time to actually do the study and get it finished. The further you are from your normal patient population, the less likely it is that you will have the time to do the research you have planned.

The best project for a busy clinician tacking a first or second try at doing research is probably going to be one that sustains motivation by answering a very real clinical question bothering the clinician. For example, the clinician may have patients with a common set of similar, but poorly defined, symptoms who are lumped into some catch-all diagnosis and who respond poorly to any intervention. They take up a disproportionate amount of the clinician's time because they can not be treated effectively and become annoying to see because of the clinician's very human response to frustration. This is the group to do research on because they are causing a problem now and the solution may help the clinician and patients directly.

First year residents and new graduate students (who have not been out in the clinical world before) frequently do not have the background in their new field to pick a project which will match quickly developing interests. It is very unlikely that the work will ever get done if the topic which grew from first burst of enthusiasm at the start of the first year quickly becomes recognized as irrelevant naivete.

2. Clinicians are often over-scheduled with patient care tasks and may have difficulty finding time to do complex projects which take them out of their clinical setting. Thus, when planing a study, time availability is the key element. Many studies can be planned so that they use the clinician's usual patient population and make a minimal impact on the clinician's time and duties. Clinical studies tend to take a long time to perform. A rule of thumb is that the more of the clinician's time that will be required to perform the study, the more funding will be required to hire somebody to either replace the clinician for performance of regular duties or to perform the research functions the clinician really wants to do.

3. The clinical research that actually gets completed - as opposed to the vastly larger group of projects which are planned and started but never finished - are those that (1) are of real interest and of clinical significance to the participating clinicians and (2) are possible to perform given the time constraints of the investigators. Burn out causes the premature demise of most projects.

4. *If you are not really interested in the topic or realize that you just do not have the time to do the study properly - don't start.*

C. Sequence of events / time line: Most clinical projects follow the same general sequence of events. With the crucial exception of the data gathering stage, the events tend to take about the same proportion of time to complete. This sequence is illustrated in Table 1 on the next page. You can follow this guide as you go though the process of planing, performing, and analyzing a study.

Table 1: Basic steps in performing an initial clinical study

Identify interesting & meaningful topic	3 topic must be capable of sustaining interest over time and with changes in experience
Take the time to think about the idea	3 narrow the topic area to something manageable and plan the actual question(s) to be answered / hypotheses to be tested
Discuss the idea with colleagues	3 nobody has all the good ideas or sees all the problems
Investigate your idea	3 perform a literature search, talk with experts outside your institution, go over records of patients similar to those you will study 3 Relate your findings to the topic!
Define the essential variables & look for confounding variables	3 the variables must be measurable! find at least two ways to measure each variable because one won't work out.
Decision time	3 Stop & think, talk with others 3 Is the idea worth pursuing or is it so limited that it is no longer interesting or practicable? If not, drop it!
Plan the study	3 include ways to take and analyze the data
Perform a very small, informal pilot trial / dry run of the procedure & techniques	3 include all data gathering instruments and methods, validate questionnaires now by giving them to a few patients of the type who will really be answering it (ask them what they think you want to know) and a few subject matter experts.
Evaluate the initial results	3 modify the study methods, instruments, subjects, etc 3 decide on group sizes using power analysis 3 recheck to insure that the data you want are actually measured by the tools you are using.
Decision time (again)	3 Do you have the resources to perform the study? 3 Are you still interested in whatever is left of the original idea? If not, stop now!

	<p>3 Does the design answer questions you are interested in? If not, stop now!</p>
Stop & think	<p>3 last look for flaws and confounding variables.</p>
Develop a formal protocol and have it reviewed	<p>3 write a formal protocol and give it to (a) an excellent, senior, clinician, and (b) a basic science researcher - between them they'll find many of the holes and provide historical and design perspectives - you don't have to use all of their critiques.</p>
Finalize the data gathering techniques	<p>3 Be sure all of the people taking the data will do everything the same way.</p>
Perform the study	<p>3 the study requires few contact hours on your part relative to the time spent planing and analyzing it.</p> <p>3 watch for problems - most studies do not work the way they are planned!!!! Don't let the study run without entering data and analyzing as you go.</p> <p>3 triple check that the data are being correctly & objectively recorded - and are meaningful.</p>
Initial descriptive analysis	<p>3 Do Not Do ANY Statistics Yet!!!</p> <p>3 graph out the raw data and look for trends</p> <p>3 think about the results and discuss them with others - there will be lots of unpredicted relationships</p>
Statistical evaluation	<p>3 apply appropriate statistical techniques</p> <p>3 it may take months to think through the results</p> <p>** are the results really clinically important???*</p> <p>3 discuss the results with others and relate them to the literature to further your understanding of what happened.</p>
Write up, present, & publish the report	<p>3 write abstract & present poster or platform talk</p> <p>3 use comments from the presentation to write a full length paper</p>
Do a follow-up when appropriate	<p>3 assess changes in quality of life</p> <p>3 follow-up needs to be long enough to determine if Rx worked</p> <p>3 defines the placebo effect rate and the failure rates.</p>

	3 be sure to track patients who were seen elsewhere.
Publish the follow-up report	3 include accurate, clinically relevant information. 3 publish somewhere that people who need the information can find it 3 be sure to include changes in quality of life

D. Quality vs. your valuable time:

1. The idea has to be superb or the project isn't worth doing: Here are some keys to doing a worthwhile project (this list is loosely based on an idea by Kahn (1994)):

a. The answer to your question **must** be truly important - e.g lead to a change in clinical practice, stop the use of a poor technique, etc.

b. The question should lead to new areas of discovery - induce others to do further research based on your findings, perhaps even point out a whole new field of endeavor, etc.

c. Ask an answerable question that is not too likely to lead you into a quagmire of uninterpretable data. This is probably going to happen to some extent but you want to minimize the probability.

d. Don't do (and redo) the same project everyone else is doing - find your own area to study which is important to you. Replication is crucial but boring.

e. Build on your previous work. It is rare indeed, that a single clinical study actually answers the question as completely as needed.

f. Balance the risk that a project will fail with the interest you have in doing it. Projects with a high risk of failure are those going furthest away from known areas because there is more chance of your measurement techniques failing, of unanticipated variables showing up, and your idea simply not working out (perhaps you really can't cure cancer with moonbeams). Although high risk projects are the most interesting and have the potential for the biggest pay off, you will get pretty discouraged if all your projects fail because of structural problems so it is wise to start with a medium risk project. As you build a portfolio of ongoing studies you can balance some really high risk ones with some safe ones so you can keep your ego, reputation, and especially funding intact.

g. Realize that most designs don't work out as you anticipate and that you are likely to have to change in the middle to succeed and that you are probably going to have to keep adding sub-studies to actually answer your question.

h. Focus on getting the job done. If you try to do too many projects, none of them will get the attention they need to work out.

2. You need to perform the project correctly or it isn't worth starting: Many, if not close to the majority, of the clinical research articles published are so flawed that their clinical value is significantly reduced or eliminated. Many report the opposite conclusions indicated by their data. Some of these are dangerous because well meaning clinicians alter their practices based on the erroneous conclusions. The vital need for good designs, even in multiple case report studies, combined with reasonable length follow-ups, can not be overstated.

a. Please don't join the group who spend such little time thinking out and planning a study that it can't answer its questions and becomes a waste of both the clinician's and patients' time.

b. Everyone recognizes that one doesn't learn to be a competent clinician in a few hours. However, experimental design and analysis is at least as complex and takes a considerable amount of time and effort to learn. As is true with any clinical profession, you get better at it as you practice it. Thus, you are not going to become an expert in experimental design and analysis in a few hours. You need to take the time to learn the skills you need to do a project correctly because (1) it's too late to fix mistakes when you are ready for data analysis and (2) you can't use the clinical literature as a guide to either experimental design or analysis by copying similar studies because so many are fatally flawed in some clinically important way!!!

E: Ready to get going? Before reading any further, take a moment to find out how much you already know about clinical research by taking the following pre-test. It will help put the rest of this book in perspective.

PRE-TEST PRE-TEST PRE-TEST PRE-TEST PRE-TEST

1. Determination of appropriateness of design:

a. Don't change your design just because you would be terrified at the idea of participating in your own study (or having your spouse or kids participate).

b. Don't worry if you have to hold off providing some of the patients with a treatment you know works. This is especially true if you know that the patient is in pain or that the treatment must be provided at some critical time after injury to insure maximal recovery. The experiment has to come first!

c. Don't worry about whether your control and experimental groups are very similar - people are people so the differences will all average out.

d. Don't worry if the control is easy to tell from the experimental condition (Eg. different tasting pills or a control treatment that even a jelly fish would realize can't actually be of any help).

*e. Don't worry about taking baseline measurements. **All** people are the*

same and you can always find some background data from some other study.

f. Don't worry about the number of subjects in each group or the number of observations. Your statistics will make it all come out fine.

g. Be sure your study has lots of variables and groups and that you collect lots of data on everything you can measure. That way you can be sure of proving something.

h. Carefully place each subject into the group you feel is best for the subject.

2. Data collection -

a. Structure your study so the subjects can tell what you want and do not bother you with confusing details.

b. Be sure to collect all of the data yourself (especially if it is very subjective data). If you can't collect it one day, just get anyone to do it any time they have a chance.

c. Be sure to keep track of which patients are in which group and whether they are getting the real thing (the one you know is best) so you can be especially careful about the recordings.

3. Data Analysis -

*a. If some of the data do not fit the rest, assume that they weren't recorded correctly and **discard** them without noting their existence in reports. Be sure to keep all of the good data that supports your idea.*

b. Since you know that your groups really are different, if you were unlucky enough to pick subjects who didn't give sufficiently different answers, add a little where it should have been. Be sure not to mention any bias that exists in your collection methods.

c. Do not worry if your data does not meet the basic assumptions of the statistical test you want to use. The computer will printout a convincing bunch of numbers.

d. Keep trying different tests until you find one that says your numbers show what you want.

END OF PRE-TEST END OF PRE-TEST END OF PRE-TEST

If you have taken any of the above statements seriously, you should rethink your understanding of research design.

Chapter 2

Defensive reading of clinical literature - does the hypothesis make sense?

A. Concept: Most clinical articles don't have a specially labeled section called "hypothesis" that jumps right out at you. Rather, you usually have to deduce the authors' hypothesis from clues scattered throughout the summary and introduction. When you pick up the article, the first thing you see is the title. A carefully thought out title should not only tell you the general hypothesis but a lot about the method used to test it and the results as well. Once you find or deduce the hypothesis, how do you know if it makes sense? It is up to the author(s) to convince you that it does. They have to provide a thorough enough introduction so you can not only predict what their question will be but to impart the belief that they know enough about the subject to ask a meaningful question and then answer it correctly. If you are left in doubt, you have no way to really judge the value of the article unless you happen to know a lot about the article's subject matter.

For example, let's say you are reading an article about the treatment of lung cancer with a new surgical technique. The title is something such as "Effectiveness of alveolar excision using laser microfusion in the treatment of lung cancer." You can pretty well guess that their hypothesis must be something like: When x-technique is used on people with lung cancer, there is some effect on their disease process. You do not know in what way it changes or how they measured the change, but you do have some idea of what they must have been trying to accomplish. Based on this assumption, there are certain bits of crucial information you would look for while reading the introduction. You would be looking for information on effectiveness of other interventions (especially surgical techniques), staging of lung cancer and survival time in relation to stage of cancer at diagnosis. If it becomes evident that this is a study on survival, you would be looking very hard for population based data on variability in years of survival. Your "alarm bell" would sound if these background points were missing because you don't know if the authors know how long a follow-up to take or how large their groups have to be for a meaningful result to be achieved. Several very famous studies using behavioral interventions for lung cancer had fatal flaws which reviewers caught by reading their introductions. Among other problems, (a) their groups were so small that known variability in survival would have made it very likely that one group would have an average survival time far longer than the other just by chance and (b) their failure to review well known life table data on survival times causes them to miss the crucial point that the apparently longer survival of several participants was not out of line with the population. Their hypotheses did not, in fact, make sense because they were too general to permit the investigators to prevent, or readers to pick-up, potentially fatal flaws.

Without a good introduction, you don't know if the question being asked is meaningful unless you are an expert. The following sections provide some guidance on how to sort through the title and introduction of an article so you can figure out whether it makes any sense.

B: An example: It's time to read sample five (the paper) at the end of this book. If you haven't read it already, please do so now so you can follow along with the logic.

C: What are the basic steps to follow for critiquing a clinical article? It is obviously imperative that you be able to read the literature defensively or you can't sort the wheat from the chaff! The question is what to look for when doing initial screenings of an article or when doing an in-depth review. In either case, it helps to have an outline of the basic things to look for in an article until you get used to critiquing on your own. Such an outline makes it easier to spot the pitfalls and weak points - and not miss a critical step which can change what appears to be a very strong article into a fatally flawed mess. Such an algorithm is presented in Section I near the end of the book. It was developed by Dr. Lori Loan of Madigan Army Medical Center (2000). Parts of it are repeated throughout this book to guide readers toward the major points to consider when evaluating a particular part of an article or protocol.

I use a much more concise set of criteria for an initial look at an article which has simply attracted my interest then switch to outlines such as those prepared by Dr. Loan if I am doing an in-depth review of it. This is my list for an initial screening:

Key elements of a credible clinical study / publication:

1. Adequate diagnosis of the subjects
2. Adequate pre treatment baseline to establish symptom variability.
3. Objective outcome measures relevant to the disorder.
4. Intensity of the intervention sufficient to produce an effect.
5. Way to check whether the intervention was successful (drug taken properly, behavioral technique successfully learned and then used).
6. Sufficient patient-subjects so result is credible.
7. Appropriate design for the question (e.g. single group, controls, believable placebo, etc.).
8. Sufficient descriptive statistics so results are clear.
9. Long enough follow-up so duration of results can be established.

This leads to the question of how much weight you can give to a technique not supported by several double-blind, placebo controlled studies. When trying to evaluate such techniques, I recommend you gather all the trials you can find which used it and look for (1) a greater effect size than would be expected from non-specific effects including an active placebo (active placebos produce greater effects than passive ones) which means at least 50% for pain studies, (2) longer duration of effects than would be expected from non-specific effects as above which, for pain studies, means at least six months, (3) observable effects on the underlying physiological cause of the pain with no change in pain if there is no change in the cause, and (4) consistency of effects across several independent studies. The point about active placebos made above refers to the finding that placebos which cause noticeable side effects also tend to cause greater magnitude of effects on the symptom being "treated" than do those placebos which do not cause a noticeable side effects (inactive placebos). This is presumably because patients don't associate

side effects with a placebo but do associate them with an actual medication. See Hammond (2002) for a review of this phenomenon and Price et al (1999) for an analysis of psychological factors which enhance the placebo effect.

D. Crucial points to look for in the introduction to an article:

1. Problem Statement:

- a. What is the problem that was studied? Is it explicitly identified?
- b. Is the problem stated precisely and clearly?
- c. Is the problem delimited in scope?
- d. Is the problem justified in light of theoretical and empirical work relevant to the topic?
- e. What does the literature say? How does the study fit with what is known?
How does it contribute to gaps in knowledge?
- f. Is the theoretical and practical significance of the problem discussed?
- g. Of what importance is the problem to medical science and practice?
- h. Do the authors provide convincing evidence that they understand the underlying physiology and epidemiology of the problem they are working on?

2. Conceptual Framework

- a. What are the major concepts guiding the study and how are they defined?
- b. Are the concepts linked to one another? How?
- c. What theoretical perspective has been used to better understand the problem?
Is a theoretical or conceptual perspective clearly identified?
- d. Does the conceptual framework accurately reflect the state of medical science?

3. Purpose:

- a. What is the purpose of the study? What was the investigator trying to find out? Do these seem congruent with the background which has gone before?
- b. What concepts or variables are specified in the purpose? What are the variables (independent/dependent)? Are the variables sensible indicators of the concepts? Do any important aspects of the concept seem to be omitted?
- c. Is the purpose logically linked to the conceptual framework?
- d. Is the purpose linked to earlier empirical work?
- e. Does the purpose precisely indicate how the study will contribute to new knowledge? For example, will the study/article contribute description of a phenomenon, explanation of a relationship between two or more concepts, or predict an outcome?

E. The next chapter will discuss how to formulate a good hypothesis so you have a better chance to avoid going off at full tilt into a project you can't get a meaningful answer from.

Chapter 3

Protocol development - formulating and maturing a question.

A. Testing the waters and thinking it through:

In chapter one, you were pummeled with the idea that if you don't have a question capable of sustaining your interest and which was meaningful to both yourself and your field, there is little point to going through all the agony and effort of getting a research project started because it probably won't get finished or be worth finishing if you do manage to drag it to some conclusion.

But, let's say you do have an idea worth pursuing. The key trick is to get from there to a doable project with an answerable question. Believe it or not, the first thing to do is just stop and think about your idea for a while. Push it around in your imagination from all sides. Try to see how you could test it within your resources. Now is the time to narrow the topic to something manageable and set out the actual questions you will try to answer as testable hypotheses. The next step is the most difficult of all because you need to risk your ego by discussing the idea with colleagues. The fact is that nobody has all the good ideas or sees all the problems. Many training programs have regular meetings at which new ideas are brought up and discussed by the entire staff. This works well if its done in a constructive, friendly manner instead of a bunch of sharks circling for the kill. Your more senior colleagues may have seen your idea tried - and failed - before so they may have crucial information that could save you from going up a blind alley or missing an obscure point which turns out to be a fatal flaw.

If the idea seems basically sound and practical to your colleagues, its time to investigate it in depth. We'll discuss the tricks to doing a background search later, but for now suffice it to say that doing a computer based literature search in ten minutes is not going to give you the information you need to determine whether your idea is doable. Not only do you need to do a literature search, but you need to discuss your idea with experts outside your institution, go over the records of patients at your institution similar to those you will work with, and basically become a true expert on the problem you are going to investigate.

B. Know what you are doing:

As discussed above, you have to ask the right question to get the information you are looking for. You can't ask the question if you don't have a solid knowledge of the field your question resides in and an in-depth knowledge of the specific part of that field impinging on your question. In other words, you need to be a true subject matter expert before you can really ask a meaningful question which you have a chance of answering. The frequency of protocols and

articles which provide overwhelming evidence that their originators haven't the foggiest idea of what they are dealing with is simply stunning. The behavioral literature is especially prone to articles in which an intervention is given for some problem which is so poorly defined that nobody could interpret the results or even tell what set of disorders might have participated. For example, the majority of the articles in which biofeedback was incorporated into behavioral interventions for back pain do not define their populations well enough to know (2) if the problems were acute or chronic or (1) what the underlying pathology was (e.g. Sherman and Arena 1992). The group sizes are usually so small and the follow-ups so short that any changes are as likely to be due to random variability as anything else. These investigators did not take the time to study the disorder well enough to do meaningful studies.

The key take home point here is that assumptions are the hidden fatal flaws waiting to destroy your study. Here are some of the major areas to look out for:

1. Know your disorder:

a. Know the expected cyclical variability and progress the disease: This is the idea that you should begin treating the common cold after about a week so your treatment will show itself to be rapidly effective. We recently completed a study on cognitive changes related to the menstrual cycle. Everyone was fine with the idea of retesting subjects for months because everyone knows that the menstrual cycle is (supposedly) cyclical. However, we were simultaneously doing the pilot for a study on the occurrence of sports related musculoskeletal injuries among female athletes. We weren't experts on women's athletics and didn't review the literature carefully enough to find out that these injuries are also related to the menstrual cycle. So, we didn't know our problem well enough to do research on it.

b. An example of the problem with not knowing the disorder recently surfaced on an internet chat line. Several people reported successfully using biofeedback to reduce the twitches related to Tourette's syndrome. The controversy that ensued related to the length of their follow-ups and the timing of treatment initiation. These twitches have apparently random, very long duration changes in baseline levels of activity during which they frequently disappear for years after being prevalent for many months. If treatment is started during a period of high activity, the activity level is likely to go down naturally and remain down for years before resurfacing. Thus, there is little way to tell if the treatment was effective using an open design.

2. Know enough to ask the question: It is very rare indeed, that anyone is expert in all the areas needed to tackle a clinical study. A physical therapist may be an expert in biofeedback for muscular rehabilitation but could not really approach a tension headache study without the knowledge of a neurologist and a psychologist. In short, a team approach is far better than trying to learn the bits you don't know because you may know so little that you don't even know that a huge morass it out there. The time to find that out is when you are planning your study - not when a journal's reviewers laugh at you.

Crucial points about working with people involved with your project:

1. Study assistants: The difference between technicians and research associates:

Technicians are paid to do repetitive tasks the same way each time while maintaining quality, etc. They aren't paid or selected to think about the effectiveness of the task or how it fits into the overall process. More often than not, thinking like a technician is deadly to research projects. Yes, many tasks have to be done exactly the same way each time. BUT, the person doing them has to understand what they are doing and why because they are closest to the task and will be the ones most likely to notice when the data being gathered have a problem. They won't know its a problem if they don't truly understand the study and their part in it. For assistants to be optimal members of the team, they need to see themselves as research associates - not technicians. They need to feel free to bring up questions and to take ownership of the study. If its partly their study, they will help insure that it works well.

The corollary to this is - don't hire people who aren't interested /able to help the project along.

2. Building the research team:

a. Include everybody - especially the people who will do the work (research associates, nurses, etc.) from planning onward. I have seen several studies killed by nurses who had to do all the work but who were never consulted during the planing stages nor included in the team during the study. In one case they were simply told to collect data at a particular time of day and in the other to give certain fluids. In both cases, the nurses immediately saw fatal flaws in the studies and could have saved them but were so upset at being ignored and treated as mindless technicians that they didn't tell the physicians running the studies until it was too late to save the studies. You don't need to have a similar situation.

b. Everybody can have good ideas and nobody has all of the answers or sees all the flaws. Truly open yourself to the team approach - not just adding names of somewhat uninterested people to the list of protocol co-investigators. It takes work to get people to actually participate and to risk critiquing your idea.

3. Know your tools and equipment: The number of people who do studies with no understanding of the instruments they use to take their measurements is simply stunning.

a. Use of the Minnesota Multiphasic Personality Inventory (MMPI) to evaluate low back pain. There are hundreds of articles showing that two of the inventory's scales are frequently elevated when people have low back pain. Unfortunately, the author's interpretations of this result was that the scales were elevated because the people had some kind of psychological problem that caused them to exaggerate their pain. There is a high correlation between how elevated these scales are and (1) the duration of back pain and (2) the rate of failure of back surgery. Surgeons were happy with the test because it is obvious that the surgery will fail

if the patient has mostly psychologically based pain. Unfortunately for this pretty picture, an orthopedic surgeon actually read the inventory and found out which items were causing the scales to be elevated. They were all items which related to very real back pain such as not being able to sit still. He published his findings in an editorial that caused hundreds of psychologists to look like total idiots because they had never read the questions on the test they were giving. The problem was that the inventory had never been intended for people with actual back pain and had never been normalized on this population. For a review of this particular problem see Sherman et al (1995).

b. "Alpha" EEG: The early biofeedback literature is rife with reports of people being trained to increase the "alpha" frequencies of their cortical EEGs. The early equipment was not able to reject electromyographic artifacts and the people doing the training frequently did not know enough to look for these artifacts or to recognize them. It later turned out that much of the apparent change in EEG frequencies was actually a change in amount of artifact recorded as subjects learned to roll their eyes up into their heads and perform other muscular actions which caused an increase in the feedback signal which was caused by an apparent increase in the amount of alpha produced when the artifact interfered with the signal. A recent example from the Internet concerned a clinician who was very proud of being able to produce more "alpha" frequencies in his EEG spectrum while meditating. As part of his clear description of the sequence of events, he noted that his eyes rolled up when producing increased alpha. None of this infers that you can't train people to change their EEGs or that the change will not result in a change in some condition. It simply means that you need to know what you are doing if you are going to use a machine to record a physiological signal.

4. Know your physiology: You can make a real fool out of yourself as well as waste your and your subjects' time and energy if you don't take the time to find out how the physiological system you are examining works.

a. Many of the early studies on blood pressure did not include repeated baseline measures. In a typical study, a patient diagnosed with essential hypertension would be brought into a clinic for treatment. A baseline reading would be taken and treatment would commence. Readings would continue as the patient returned for subsequent visits during the treatment period and for a few weeks after the end of treatment. Interestingly, nearly all treatments were effective. Of course, we now know that the subjects' blood pressures were initially elevated by stress responses to the new environment and gradually decreased as the subjects adapted. In our early studies on biofeedback and relaxation training for essential hypertension, we brought the controls back as often as those subjects receiving therapy - and the progressive decreases in blood pressure were fairly similar.

b. Fingertip temperature: How many kids are out there who are under the impression they can raise their fingertip temperature five degrees in five minutes? Plenty. This is because teachers frequently get hold of temperature biofeedback machines, attach the fingers of nervous students to the device and tell them to raise their fingertip temperatures. The kids vasoconstrict as they are being hooked up to the device and relax when it doesn't kill them so their hand temperatures go way up. For the very few cases in which the kid gets a second chance, everyone is really puzzled when they can't raise their temperatures much if at all - and then it

goes down. This would just be an amusing story if therapists weren't failing to take adequate baselines in their practices (Sherman and Heath 1998).

5. Know your intervention: It is quite obvious that you are not likely to show an intervention to be effective if you give too low a dose of the test medication or put in the wrong size implant. Everyone recognizes the need to provide for a leaning cure when testing a new surgical procedure. But what about behavioral interventions done by undertrained people who have little idea what they are doing? This gives the same effect as using too low a dose of a drug being tested. In other words, a behavioral intervention is as likely to fail in the research setting as in the clinical setting if it is poorly performed. So, if you are going to test the effectiveness of a behavioral intervention, you have to set up at least the same level of quality control measures as you would for drugs and surgeries.

6. Get the logic straight on what factors are being compared with each other in your study design: You are going to be even more confused by your results than you might be otherwise if you have a major flaw in the logic of your prediction. For example, in the early extrasensory perception (ESP) tests people were seated in front of an experimenter who was had a deck of cards. The well shuffled deck contained an even distribution of four (or five) different types of cards - each with a distinct shape on it. The experimenter would pick up one card at a time and look at it. The subject was supposed to guess which of the four shapes was on the card. As there were four shapes, the subject has a 25% chance of guessing the correct shape. When the deck had been exhausted, the experimenter would add up the number of correct guesses. If it was "significantly" above 25% the subject was said to have exhibited ESP. The experimenters were consistently amazed when their best subjects were retested and didn't show any ESP - they performed randomly. Some even did much worse than random. This was interpreted as "blocking" their ESP because of increased pressure to do well. This is obvious nonsense as the investigators ignored the crucial point that they were comparing the results of many subjects performing the same task. The bell curve of random responses assures that a few people will guess more cards correctly than random and a few will do less than random. If you have a large enough deck and calculate how well a person is picking after each fifty cards or so, sometimes the person will be above random for one of the sequences. The point is that occasionally one person will be out on the bell curve by chance. If you test enough subjects and have a small enough deck of cards, eventually somebody should get them all correct just by chance. In fact, it is surprising that the ESP researchers didn't report on anyone getting them all right. Of course, when a person was retested, they were very likely to achieve a chance score again (statistically called "regressing to the mean") so didn't demonstrate ESP on the second trial. By chance a very few of those who did above chance on the first trial did worse on the second. I do not know if the number of people tested relative to the number who achieved above or below chance level fits a normal bell curve because I have never been able to find the data. Given the hundreds of people tested for each identified as having ESP, the ratio feels about right.

This example does not imply that ESP does not exist. It simply points out that it can't be proven with a botched study and that great care has to be taken to insure that the design insures that intersubject comparisons (as should have been done) and intrasubject comparisons (as they did) are done when appropriate.

C. Back to the hypothesis again: Your hypothesis has to be very clear to help you hone in on just what you want to know. Table 2 presents examples of weak and strong hypothesis asking the same question. A weak hypothesis is one that doesn't ask its question in such a way that it can be readily tested and answered.

Table 2

Strength and testability of hypotheses:
Weak and strong hypotheses asking the same question

(Table based on an idea I first saw used by Silverman (1977) and repeated by others dozens of times since.)

Weak / not directly testable / less answerable	Strong / directly testable / more answerable
That biofeedback is effective for treating urinary incontinence.	That urethral sEMG biofeedback given daily for two weeks is at least 50% more effective in reducing both number and extent of urine leaks than Kegal exercises practiced daily for two weeks among otherwise healthy female athletes between the ages of 21 and 35 diagnosed with exercise induced urinary incontinence.
Determination of which of six external fixators is easiest to apply.	Determination of which of the six external fixators supplied by the US Army is rated by Army orthopedic residents as easiest to apply to compound fractures of the long bones under combat conditions.
That relaxation training increases the rate of wound healing.	That subjects given autogenic exercises emphasizing warming the hands given daily for three weeks before elective hand surgery show primary wound closure at least 25% sooner than age and problem matched controls receiving body awareness training exercises for the same amount of time with the same therapist contact.
That exposure to pulsing electromagnetic fields increases the rate at which stress fractures heal.	That exposing male soldiers in cadre status on reduced activity profiles who have radiographically demonstrated tibial stress fractures stable for between four and six months to 475 watt, square wave electromagnetic fields pulsing 7,000 times per second daily for four weeks will (1) result in at least 30% healing upon bone scan at the end of the exposure period and (2) the recovery curve of treated subjects (showing complete healing upon bone scan) will be at least 25% faster than the curve produced by historical controls from this population.

Sample Study

Take a look at the hypotheses for samples one and two:

1. The exploratory pilot's hypothesis:

That application of PEMF to the inner thighs of migraine headache sufferers will result in decreased headache activity.

2. The full study's hypotheses:

a. That exposure of patients with chronic classic migraines to PEMF over the inner thighs for two weeks will result in at least a 50% decrease in frequency of migraine headaches for one month for at least 80% of the subjects exposed to actual PEMFs and that this difference will be statistically greater than the reduction shown by those exposed to placebo PEMF.

b. That the above reduction can be followed for up to six months to determine the slope of return to baseline levels of headache activity.

The hypotheses are quite different. The first is an exploratory hypothesis which is not particularly sharp because the investigators had little idea of what to expect or how their intervention was likely to work. By the time they were ready for the second study, they had a pretty good idea of what they wanted to find out so they could limit their question - and the design - to a very focused point.

D. Complexity vs. your own resources and knowledge:

You need to know when you have the knowledge and resources to perform the study on your own. It is especially crucial to build the right team for the job because there are just too many weak points you won't even know exist without the correct expertise.

1. Within your means and experience: If you want to do a simple study comparing number of days of hospitalization for two techniques used to fix the same problem, you can probably tackle it yourself if you have contact with enough physicians or surgeons doing the two techniques and can match or randomize the patients.

2. Reaching: If you want to find out the long term effectiveness of a technique in common use but don't have enough subjects to do it yourself, life gets complex in a hurry. You will wind up having to enlist numerous clinicians from many sites who do the technique similarly. You will have to control for all sorts of patient and technique variables which can sink your study. You will need paid clerical support to do this right.

3. Fools rush in: If you want to determine the relative cost/effectiveness of several techniques, the best thing to do is not do it unless you are an expert at research and have funds to hire plenty of help. These studies are huge and complex.

Sample Study

Look at the team created to perform the pilot study in sample 1.

The originators of the idea had expertise working with pulsing electromagnetic field generators and with tension headache patients but no experience working with migraines or medications.

The knew that there are incredibly complex migraine variants and that the huge variety of drugs given to prevent and treat migraines have numerous effects on the vascular system. Thus, a neurologist had to be added to the team to insure that a huge area wasn't missed.

E. Choose the right outcome measures:

Your design and your outcome measures have to be exceedingly carefully fit to what you want to find out. For example, orthopedic outcome studies aimed at determining the effectiveness of replacing a bone or joint frequently emphasize what happened to the hardware and leave out what happened to the patient (Garland 1988). If you want to determine how well a knee replacement technique worked, be sure to include how well the patient is functioning in the normal environment, how much pain there is, etc. Information on how well the hardware is holding up is important but may not be the main issue depending on what you want to know.

There is nothing as useless or embarrassing as choosing an outcome variable which does not reflect the status of or can not reflect changes in the problem you are studying. This is why you must not only be an expert in the disorder you are studying, but you must also take the time to truly understand the outcome measures you propose to use and the instruments used to perform the measurements. It is truly amazing how often a study is proposed in which the main outcome measurement will be made by a questionnaire the investigators made up from scratch but have no plan for validating or, worse yet, plan to use a well validated, standardized questionnaire which clearly doesn't measure the variable they are interested in.

Incorrectly used analog – visual pain scales are one of the worst problems because users rarely bother to review the literature on how to use them. For example, the typical 0 – 10 pain rating scale is useless unless 10 is defined as something such as “so much pain I would faint if I had to bear it for one more second”. Without that top limit, people keep changing what they mean by a ten as they experience greater pain. For more information about pain scales, see protocol “sample 1”. For details on how to assess and treat patients with chronic pain while performing clinical research, see Sherman (2004), Pain Assessment and Intervention from a Psychophysiological Perspective.

Biofeedback studies frequently use measures of change in the syndrome being treated as an indication of whether the skills being taught have been learned. This is a real problem because many studies show biofeedback techniques to be less effective than they probably are because

they include people who did not learn the technique in the first place as treatment failures. Instead, the treatment outcome should be related to success in learning the technique so people who did learn the technique can be separated from those who did not. For example, a study might be performed to determine whether teaching patients with musculoskeletal upper back pain to reduce their levels of bilateral trapezius muscle tension using biofeedback would result in lower levels of pain. Trapezius muscle tension would be recorded using standard surface electromyographic techniques and the subjects would be shown instantaneous changes in their average muscle tension on a computer display (biofeedback). Let's say that fifty-one subjects who met the entrance criteria completed the study's follow-up period. If thirty of the subjects showed no improvement while twenty-one met the criteria for improvement, the treatment was not particularly effective. Or was it? The actual contingencies in such a study are shown in Table 3 below. In fact, the investigators were not particularly successful at teaching people the technique because slightly less than half met their criteria for demonstrating an ability to reliably reduce their muscle tension and showing a 20 percent decrease in trapezius muscle tension from before to after training. The investigators set the learning criteria at a level at which they felt that people who could control and reduce their muscle tension to that extent should show a decrease in pain. Of those who did learn the technique sufficiently well to meet that criteria (as measured by surface electromyographic levels recorded during the last training session and at the end of the follow-up period), three times as many subjects showed improvement as did not. Thus, the technique actually was effective for those who learned it.

Table 3

**Effect of using the correct outcome measures
on understanding the results of a clinical study involving skill acquisition**

	Subjects who learned the skill	Subjects who did not learn the skill
Subjects who showed no improvement = 30	6	24
Subjects who improved = 21	18	3

Not all outcome variables actually measure the disorder you are treating. Some variables measure factors which influence either the disorder's severity or accuracy of the patients' reports about the disorders' severity. For example, our perceived intensity of pain is highly effected by current levels of anxiety so changes in anxiety level need to be monitored so the magnification of pain reports due to increased anxiety can be factored out. Failure to perform measures which help explain extensive variability can result in so much unexplained variability that any effect of the treatment is submerged in a sea of random changes.

A particularly deadly combination in the misuse of outcome measures in clinical studies occurs when they are not made objectively and not taken for long enough after the end of treatment to be meaningful. This is frequently seen when treatment success is evaluated through the sole technique of the treatment provider asking patients whether they feel that they have improved as they walk out the door after their last therapy session. Patients certainly aren't going to be very accurate in their report and, even if they were feeling better for the moment, there is no indication of how long they will stay that way. Yet this kind of result frequently appears in manuscripts submitted for publication - and all too often - in print. These problems will be discussed in greater detail in the study design sections of the book.

Sample Study

Look at the outcome measures chosen for sample study 1 (on headaches)

Can the investigators tell whether the problem they are investigating has changed during the study?

Are the outcome measures appropriate to the problem?

Are there other outcome measures they might have considered to help explain variability in response to the intervention?

F. Decision time again:

By the time you whittle down your idea to a doable, testable hypothesis or two, you may realize that the part of your idea which you can test is so limited that you just don't care about the answer. This frequently happens when you don't have any funds to perform a project. If you find yourself in this predicament, you might want to drop the project before you get started on something that will take more effort than it is worth to you and which you are likely to drop in the middle anyway.

The problem with dropping a project after you have discussed it with half the people at your institution and have spent months investigating and poking at it is that you have grown attached to it and everyone now recognizes that you are the local expert on it. Your colleagues are probably sending every patient with even second cousins to your problem to you and are always stopping you in the hall to find out how your project came out. In other words, the old saying that it's harder to get out of something than to get into it holds as true for clinical research as for most other things.

G. Summary:

Table four contains a summary of many of the problems with clinical research questions and a few potential solutions.

Table 4

**Common problems with clinical research questions
and potential solutions to the problems**
(This table is heavily modified from one in Hully and Cummings (1988))

Problem with the question	Potential solutions
Question is vague, not testable in current format	Think out exactly what it is that you want to know and write the hypothesis to answer only that specific question.
Question is too broad to tackle in one study	<ol style="list-style-type: none">1. Decide which part of the question is most crucial to giving you the information you need and answer that question.2. Perform a series of related studies which can answer each specific question in turn.
Insufficient subjects available to answer the question (too much variability, etc.)	<ol style="list-style-type: none">1. Find ways to limit the variability in response to the intervention (e.g. (a) increase the power of the intervention, (b) decrease the variability in the test measurement system).2. Find sources for more subjects (e.g. multicenter study, better advertising)3. Expand the inclusion criteria and / or narrow the exclusion criteria if it can be done without destroying your study.
Methods beyond your skills	<ol style="list-style-type: none">1. Form a team so all the required skills as well as the broad base of knowledge underlying the skills needed to recognize when something is going wrong or missing are available to you.2. Find an alternative method which is just as good but which you can do on your own.3. Learn the skills.
Too expensive	<ol style="list-style-type: none">1. Find a less expensive way to do the tests and measurements (e.g. add colleagues who

	<p>might do the laboratory tests for free or at cost).</p> <p>2. Choose different tests or a different experimental design which might not be the best available but sufficiently accurate to answer your question without using too many more subjects.</p> <p>3. Think about which organizations (manufacturers, insurance companies, etc.) might benefit from your results and ask them to help with the cost by loaning you equipment, paying for tests, etc.</p> <p>4. Apply for a grant.</p>
Contains elements which are potentially unethical	<p>1. Consult with your institution's ethics counselor, your chief, and / or the chief of the institution's Human Use Committee.</p> <p>2. Change the design so the question can be answered ethically. Some questions simply can not be answered directly and ethically.</p>
Question is clearly not novel	Phrase the question as a purposeful attempt at replication for some specific reason.

Chapter 4

Background and literature searches

As discussed in the preceding chapter, you need to know what you are doing if you are going to get the study done correctly. Of course, you need to know what similar studies have been done. But, you also need to do a thorough investigation on the techniques you propose to use to achieve your goals. When I say "investigate" I do not mean sitting down at your personal computer and doing a ten minute search of MEDLINE over the Internet. Rather, I am referring to a thorough, exacting investigation comparable one the FBI might do when going after a spy.

While it is true that computer data bases have gotten better over the years, they still miss crucial articles regardless of how sophisticated the search and searcher are. The odds are that you are not an expert in searching each of the dozens of data bases which may contain crucial information needed to conduct your study properly. The data bases do not overlap on significant areas and each has its own idiosyncrasies which make it difficult to search. Even professional reference librarians can not do an entire background evaluation solely through the Internet's databases at this time. People who count on a literature search to find all the articles they need are usually sadly disappointed. Occasionally, the search misses the mark so badly that nothing turns up and the investigator is left with the idea that nothing is out there on the topic of interest. A good way to make a mistake in your study is to stop with a computer search. You need to go further.

This does not mean that you shouldn't use your computer and modem to perform the initial data searches. Please do. In fact, unless you are very lucky, there will not be a professional medical reference librarian at your side every time you want to do a search. So, you need to learn how to use the Internet to find and search the National Library of Medicine's MEDLINE data base and other sources for the information you need. The easiest way to get to Medline is through NIH's "PUBMED" site which is WWW.NCBI.NLM.NIH.GOV/PUBMED. Given how quickly the Internet changes, you may have to do some searching on your own to locate the site by using search engines such as google.

Once you find an article closely related to your topic, use the citation search index to see who quoted it. Look through the reference section of all the articles even close to the topic to find better articles. Be especially sure you read the basic articles on your topic - don't count on somebody else's interpretation of what they said! People frequently simply copy somebody else's rendition of what a previous article said - and get it wrong. If you do this, you could be counting on somebody who got it backwards. The best example of this I know of is the dozens of articles on the MMPI's conversion "V" who all quote an article by the originator of the concept. There is nothing in that article about the conversion "V". This is the only article the originator published at the time and the title could lead one to believe that the information is, indeed,

contained in the article. However, the information is in the author's PhD dissertation rather than the article. Obviously, somebody made the assumption that the information was in there and quoted the article without reading it. Everybody else just copied the first person (and didn't read the article) so the mistake was perpetuated. This was a major factor in the clinical misuse of the MMPI which resulted in numerous patients being incorrectly classified as being hypochondrical! Real clinical harm was done to thousands of real people because of poor literature reviews.

Eventually your search will turn up one or two investigators who seem to be experts at the disorder or technique you are investigating. They turn up as first or second authors on numerous related papers over a period of years. These are the people who are more likely than anyone to know what has been tried and failed - but not published because it was a flop. Of course, they will also know the tricks of the trade to getting a particular piece of equipment or technique to work. The literature is one to three years out of date. These key people will know what studies are going on now. Thus, if you want to save yourself reinventing the wheel, it is an excellent idea to contact one of these folk.

Are you thinking something like:

"What's that?! I'm supposed to call some world expert who never heard of me?? Even if I could get past his secretary, he'd just hang up." Or: "I'm so ignorant *she'll laugh* at me - and I'll never be able to talk to anyone again (snivel)." "When I tell him my idea he's going to steal it and run it in his great big lab stuffed with graduate students and money before I can even get it off the ground!!"

Take heart! Most "experts" are simply folk who have been working on a technique or disorder for a while. This is probably their main professional interest and they probably love to talk about it. Mentoring is great for the ego and I know very few people who won't unbend enough to help anyone who asks in a reasonably friendly (read that as *unchallenging*) manner. They may make you pay for their advice by bending your ear for extra minutes (hours?) with more information and stories than you really want to hear. But, at least you will be up to speed on what works and what doesn't.

Will she steal your idea? Reports of this are vanishingly rare in surgical and behavioral research but terribly common in the basic sciences and biochemical/medical arenas. I am under the unsubstantiated impression that ideas tend to be stolen more by anonymous manuscript reviewers and people sitting on grant review panels than by people you actually talk to. I have called dozens of senior experts over the years and have never had anything but a positive response. Some haven't been too cordial or helpful, but they talked. Nobody ever stole one of my ideas but that could be a factor of how (un)interesting they were. On the positive side, I am aware of numerous instances where an initial conversation lead to collaboration on the proposed project and, occasionally, to long term professional interactions and friendships.

The next approach to background investigation is to go over the records of patients who have the problem on which you want to work. This will give you a good idea of how much information is already available and how many of them are actually around. People tend to vastly overestimate how many patients of any particular ilk are available. This is especially true of the ones with chronic conditions who you do not like to work with as they keep coming back. You

can even talk with some of them. Tell them what you want to do. Unless you are doing a particularly obscure study on genetic markers or something, you may be very surprised at how sophisticated their insights can be. This is especially true when it comes to establishing outcome measures.

Once you have completed your background search you still need to do a few things before you are ready to get started. First relate the results of your search to your idea. It is amazing how many introductions to articles do not support or really relate to what the authors actually did. Second, use the information you have gleaned to try out your techniques. The odds of their working the first time are fairly low even if you know what you are doing with them. You may have to go back to your sources for a few more tidbits before you are done. Don't be surprised if you have to review them again once your results are in because results rarely take an expected form or direction. In fact, one way to tell an article isn't faked is when at least some of the data are scattered all over and a few bits don't make any sense at all - even in a properly designed and apparently well executed study.

Chapter 5

Determining feasibility

This is a very short but crucial chapter. In fact, I made it a separate chapter to draw your attention to the problem.

If you do not have the resources to perform the stage of the project you are up to (be they money, patients, time, collaborators, equipment, etc.), you should not start and hope they will magically appear. This happens most frequently with not having enough time and patients.

A. Your time: Before tackling a major project you should get an idea of how much time it will really take you and your colleagues to perform the study. This is usually done through conducting a pilot study or practice runs. Seemingly innocent parts of a study can suddenly mushroom into gargantuan time sponges. Be especially wary of conducting interviews or making observations of real people doing real activities.

B. Subjects are limited resources: Having enough subjects to do a study is always a sore point. Somehow there are always vastly less people than expected who you can actually find that meet the entrance criteria and are available to participate. The medical center at which I am doing a tibial stress fracture study has a computer based system which tracks every patient that walks through the doors and records the diagnosis, treatment, and everything else anyone would ever want to know about that patient. I know how many people have come in over the last two years with stress fractures. I even know their dispersal pattern across the months and years. We looked up enough records to be absolutely certain that sufficient patients are in the system who have severe enough stress fractures of the type we need and that they are in the area long enough to participate. So, here we are in the middle of the study and guess what?

C. Money: Cash is a truly limiting problem in research. You may have a study you can do absolutely for free because you have the time, equipment, supplies, and subjects. If you do, you are very lucky. This is quite rare because even if you have all of the above, your organization may demand overhead for the space and resources used to house the study. They may even want compensation for your time. Usually you will need at least some funds to pay for minor supplies, photocopying, and the like. Most organizations include them in the cost of being there - but some don't. We'll go over sources of funding in the chapter on grants.

Lack of money turned out to be an unanticipated factor in destroying the feasibility of a study was one I tried to perform on the efficacy of biofeedback for low back pain. The idea was to have clinicians using biofeedback for low back pain tell us what their usual diagnostic and treatment regime was, fill out a form on each client indicating how the client had done during therapy, and to have them give their clients logs of their pain (and associated variables) which would be filled out by the clients before, just after, and a few months after the intervention. The logs were to be mailed directly back to us in pre-stamped envelopes. We had enough money from a tiny NIH pilot grant to pay a research associate to send out and receive the information. It

never occurred to us that clinicians would not participate unless they were paid for what appeared to us as being a minuscule amount of time required to fill out the forms and to give their clients a packet of logs while explaining the study to the clients. No private practice clinicians participated. The exit survey showed that they didn't participate because they were not paid.

Sample Study

Look at the budget for the grant in sample 6.

Notice that there is no funding for released time for the investigators. If they are going to do the project, it is going to come out of time they could be earning money seeing patients.

Notice that there are no funds to pay the patients for participating. The patients have to make a lot of trips to the hospital for free. They will have to be quite certain they are going to get something out of it or they aren't going to participate.

Notice that there are no funds to repair the equipment. If they or the university can't pay for a broken piece of equipment, what happens to the study?

Notice that there is only a set amount to pay the technician. What happens if the study takes more hours to run than they anticipate?

D. The experimental design: Some designs are deceptively simple. When you try to run them, they just don't work. For example, we recently completed a study which required people to be treated for half an hour at a time. We knew that we couldn't get people on and off the machine instantly so left 45 minutes for each appointment. We scheduled the appointments back to back throughout the day because we had a limited amount of time to get the study done. Of course, people randomly came just a little late or a machine broke down. We never caught up with our schedule. This was a very minor complication compared with most nightmares which destroy studies. The typical ones are when a test takes twice as long to run as anticipated or turns out not to be practical at all. Even something as simple as having to send samples to a distant lab can turn out to be a show stopper when there just isn't an economical way to collect the samples properly and get them out in an appropriately timely manner.

E. Try it first to make sure everything works: There is no substitute for trying every aspect of your study out in advance to find the weak points. Be sure you include trying every device that you will use. It is amazing how many simple looking gizmos simply don't work as advertised or produce the precision of data you need. This will be discussed in greater detail in the chapter on logic of designs.

Chapter 6

Research ethics

A. Overview: Even the best intentioned clinical researchers frequently do not initially recognize the ethical problems and quandaries which can result from having sick people, who came to an institution for care, participate in studies. There have been so many problems over the years that a huge body of literature on clinical ethics has developed. The horrors of the clinical studies conducted on unwilling prisoners by the Japanese and Germans during World War Two gave considerable impetus to development of both Federal regulations and international treaties. The infamous "Tuskegee" syphilis study is an example of both changes in acceptable procedures and the need to continue monitoring studies as they progress. In this study, poor black males with syphilis began participation in a legitimate treatment study which was never stopped after better treatments for this disorder became available.

There are many excellent books on research ethics. They help investigators and Human Use Committee members become aware of complex ethical situations which exist and the rules governing research. The problem of conflict of interest is especially crucial as practitioners frequently confuse their roles as therapists with their roles as researchers and wind up in ethically compromising situations. Practitioners with an economic tie to the sponsor (e.g. own stock in the company) frequently take more risks with their subjects than do non-invested investigators. This information is delineated in Soece et al's (1996) book on the subject. Monagle and Thomasma's 1994 book on health care ethics and Levine's book on ethics and regulation of clinic research give excellent delineations of the arena.

The rules we work under are codified in the Nuremberg Code of 1949, the 1964 treaty entitled "World Medical Association Declaration of Helsinki" (both presented at the end of this chapter) and the National Research Action (Public Law 93-348). The regulations for investigators and human use / institutional review boards are codified in the HHS Policy for the Protection of Human Research Subjects (published in July, 1981) and the Belmont Report (1988). The combination of treaties and Federal laws place specific responsibilities on both the investigator and the sponsoring institution.

It is your institution's responsibility to help you understand them, insure that your protocol meets current ethical standards before it begins, keep you abreast of changes in them as your research continues, and monitor your study for continued compliance. No human research can take place without being monitored by some outside organization. Investigators are not permitted to decide on how risky a study is. In fact they are not permitted to use themselves as subjects without outside approval because of the loss of objectivity resulting from emotional ownership of the idea and the study.

As an investigator, it is your responsibility to be familiar with the laws and treaties governing research and to follow their guidance. So, what do these regulations want you to do? The bottom line is only to conduct studies in which you would be a willing subject or in which

you would permit a loved one to participate. This means that under no circumstances shall the welfare or treatment of a subject be put second to participation in a research study. No patient's treatment can be substantially delayed or compromised to a clinically important extent as a result of participation in a study. It is the absolute responsibility of the investigator to protect participating subjects from risk. You MUST tell the subject everything relevant about the study and its risks. If they are randomized, you must explain what that means and the effect on their therapy.

All of the ethical guidelines for working with patients apply to patients participating in clinical studies. The clinical "duty to protect" specifically extends to research subjects. Thus, if an investigator becomes aware of a risk to someone not involved in the study because of actions or potential actions on the part of a study participant, the investigator must take appropriate steps to remedy the situation. All of the limitations on personal relationships between patients and therapists continue when they are in the roles of subject and researcher.

In the United States, treatments must be both safe and effective to be accepted while in Europe, the emphasis is on safety with the investigator left to determine efficacy (Fuson et al 1997).

Many top of the line journals such as the Journal of Bone and Joint Surgery (Einhorn et al, 1997, Fuson et al 1997) will not accept articles which they feel contravene any of the ethical principals discussed in this chapter and require all authors to assure the journal that they have not violated ethical principals. Most professional societies (such as the AMA and the APA) have very specific ethical guidelines which include research activities with both human and non-human subjects.

Monagle and Thomasma (1994) combined the ethical principles delineated in the Belmont Report (1988) and Beauchamp and Childress's work to produce the following summary of ethical principles research:

1. Respect for Persons: The duty to respect the self-determination and choices of autonomous persons, as well as to protect persons with diminished autonomy, e.g., young children, mentally retarded persons, and those with other mental impairments.
2. Beneficence: The obligation to secure the well-being of persons by acting positively on their behalf and, moreover, to maximize the benefits that can be attained.
3. Nonmaleficence: The obligation to minimize harm to persons and, wherever possible, to remove causes of harm altogether.
4. Proportionality: The duty, when taking actions involving risks of harm, to so balance risks and benefits that actions have the greatest chance to result in the least harm and the most benefit to persons directly involved.
5. Justice: The obligation to distribute benefits and burdens fairly, to treat equals equally, and to give reasons for differential treatment based on widely accepted criteria for just ways to distribute benefits and burdens.

B. Getting patients' agreement to participate in a study: True voluntary, informed consent is incredibly difficult to get. Two major issues (and a host of minor ones) are involved: (1) coercion vs. free choice to participate and (2) actual informed consent. The rules for informed consent are

specified for all research performed in the United States by the Belmont report (1988).

1. Voluntary consent: When patients are asked to participate in a study, the person asking them to participate is usually their health care provider. This puts the patient in a real bind. Patients come to health care providers to get better; Not to participate in research studies. From the patient's point of view, declining to participate in their provider's study could damage their relationship with the provider and reduce the chances that they will get optimal care. The bottom line is whether a critically ill patient will turn down the potentially lifesaving provider? The clear answer is "no". Thus, consent is rarely voluntary if the provider is perceived as being involved in the study.

The problem is accentuated if the patient being solicited is not mentally and/or legally competent to consent to being in a study. For example, a four year old is not likely to understand an explanation of a double-blind, cross-over, placebo controlled study on psychologically based stomach aches. He just wants his stomach aches to go away. His parents need to consent to his being in the study. In this example, the child can not make an informed judgement so the parents will make the entire decision. But, what if the patient is a normal ten year old. By this age, the child can understand the basic idea of the study. You **MUST** obtain the child's assent to participate if they can understand the study. If the child says "no" and the parents say "yes", the child can not participate.

If the patient is so mentally disabled that a rational decision can not be made or if the patient is unconscious during the crucial time consent must be given (as when testing trauma devices in life threatening situations), the next of kin or a judge must give consent.

Prisoners of any type (criminal, political, military, etc.) can not be used in clinical studies because they can not give true voluntary consent without fear of retribution from their guards and similar problems. They can participate in studies which are evaluating the prison process, health issues of prisoners, etc. but getting approval can be torturous.

Most writers now feel that it is not appropriate to pay subjects to participate in a study so much that the reward would influence their decision to participate. The amount subjects get to participate should balance their inconvenience and costs for participation such as extra trips to the clinic.

(2) Actual informed consent: The consent form in the first two sample protocols at the end of this book are typical of the very simple consent forms that go along with very simple studies. As studies become more complex and have more risks, the consent forms become lengthy documents (eight single spaced pages is not uncommon) where dozens of risks of varying degrees of likelihood are detailed. The odds of a well educated non-clinician being able to understand one of these, even if given days to look it over are very low, regardless of how clearly it is written. The odds of a typical patient given a few minute verbal explanation and handed a consent form to sign while waiting for the provider to return actually comprehending the form in a useful way is nearly nil. Numerous studies have shown that most subjects have little understanding of the studies they are involved in once the study has reached the placebo group stage of complexity. For example, Park and Covi (as quoted in Spence et al 1996) found that, even after informed consent, about half of the adult patients receiving the placebo in a non-blind, non-randomized study believed that they were definitely receiving active medication. A neutral party, such as a minimally involved technician, should take the time to go through the consent form with the subject and insure that all questions are answered. Every attempt should be

made to permit subjects to take the consent form home and read it at their leisure. This gives the subject an opportunity to discuss participation with family members and respected members of the subject's community. This can be very important because many subjects are unwilling to disclose their need to take the form home. For example, various religions have differing prohibitions against various procedures occasionally found in research projects and the subject may wish to consult with a religious leader.

In group settings when many subjects are solicited at once, no subject should be consented unless that person has an opportunity to ask questions privately and to decline without the other members of the group being aware of the choice.

C. Fraud - faking and fabrication of data - when you don't get the results you expect: Fraud in clinical and basic research appears to be distressingly common. Investigators appear to totally fabricate or "enhance" their data in the direction they would like it to go in with considerable frequency. The most common types of fraud in reporting research data appear to be (1) entirely fabricating the results of a study which was never done at all, (2) selectively reporting data to eliminate the readings / subjects that don't go along with the direction you want, and (3) trimming and outright changing some of the numbers so they come out the way you want them to. This is a very different situation than the unconscious twisting of subjective data to which everyone can be prone. Rather, it is outright fraud and can be prosecuted under the law. According to the FDA's and NIH's chart audits, it occurs all too frequently. A fourth type of dishonesty consists of pretending that other work supports yours even though it doesn't or failing to read the other work and then quoting it. Broad and Wade (1982) have extensively reviewed these problems and their prevalence and Levine (1986) discussed them in relationship with broader ethical issues in clinical research. The rules governing misconduct in research in the United States include 42 CFR 50.102/3 and 45 CFR 689.1 with prosecution under 18 USC's 1001 and 1341. The House Subcommittee on Oversight and Investigations conducts investigations into scientific fraud.

How common is scientific fraud? Dingell (1993) reports the results of a study showing that about 40 percent of graduate school deans knew of confirmed cases of scientific fraud occurring at their institutions within the last five years. He also reported the results of a large survey of scientists which found that 27% had personally encountered at least two instances of research they suspected was falsified, fabricated, or plagiarized within the last ten years.

1. Entirely fabricating the results of a study which was never done at all: The most famous example I know of this kind of fraud occurred with a journal dedicated to work on twins who had been separated at birth. After years of publication, someone realized that more reports on separated twins had come out than there could be pairs of separated twins. An investigation led to the finding that the editor had made up many of the articles himself entirely from his imagination. Dingell (1993) reports several such cases.

2. Fabricating results of a study that didn't work out: A famous example of this genre involved a model for tracking the effectiveness of treatments for skin cancer. A model was developed in which cancerous skin on black haired mice was transplanted and then grown on white haired mice. Changes in the size of the skin cancer could be seen easily as the size of the black area changed with treatment. This model was well accepted in the scientific community and numerous publications from many first rate universities used it to demonstrate the

effectiveness of treatments over a period of many years. A graduate student from a small university who wanted to use the model kept trying to transplant the skin without any luck. He called the originating university numerous times and eventually arranged a visit at his own expense. He was unable to get the skin grafts to take even under close supervision at the host university. While there he made friends with a laboratory technician who eventually showed him what was being done. It turned out that the graft techniques didn't work at all. However, a black magic marker did. All the student investigators had to do was imprint several virtually identical mice with the same ear code. The students subsequently drew the desired size spot on a sequence of identically marked mice so the senior investigators would think the study was working. The graduate student told the authorities and the scandal hit very hard. It later turned out that student investigators at several schools had independently developed the magic marker technique when they couldn't get the grafts to stick. The original perpetrator was identified and literally forced out of the country. He was a native of a different country and didn't quite have US citizenship when the crime was discovered. Dozens of papers were recalled, numerous mid level investigators were dismissed from their academic positions and several very high ranking, senior investigators lost their grants and were prohibited from doing research for many years. Their schools were ordered to return the funds but I do not know if they actually did so. The basic problem was that the senior investigators oversaw planning of the studies but had nothing to do with the work at all. The intermediate level investigators left all of the work to the students and only helped plan, analyze, and write up the studies. Nobody checked the work of the medical students who were under tremendous pressure to produce positive results. Nightmares such as these forced a Nobel laureate to resign as president of a major university and retract numerous articles (Dingell 1993).

3. Selectively reporting data to eliminate the readings / subjects that don't go along with the direction you want: This is done by vast numbers of investigators who decide that an unwanted number is an "outlier" without performing the appropriate tests to show that the number is, in fact, so far from the possible values that it should be excluded. An example would be a retrospective study on human six month olds' weights and finding a recent record of one weighing 900 pounds. This is unlikely enough so it can be discarded. But what if the child's weight was recorded as being 20 pounds? Investigators and their technicians tend to make snap decisions when they see a value which seems to be out of scale with the others. You can always find an excuse to drop a patient from a study if you try hard enough. You can also convince yourself that a particular rat was acting strange or that you contaminated a particular test-tube.

A famous test of people's tendency to rid themselves of ostensible outliers in order to get results they think are correct was done in a large elementary school with many classes for the same age grade of students. All of the teachers for that grade were individually told to measure the heights of their students and that their class's average height would be reported as the average for the school. None knew that all of the other teachers were also measuring their students. Of course, young kids vary tremendously in height. Each teacher had one or two kids who were obviously much taller or shorter than average. Including them would cause their results to be very different than what the average height of all the kids in that grade probably was. Virtually all of the teachers eliminated the outliers and, of course, came up with the wrong means. When the data for all the classes were combined, the data collected by the teachers was not representative of the actual average height of the students in that grade.

The moral of the story is - don't assume that you know what the values should be! The

best approach is to report all of the values and present analyses of the data both with and without them. This is vital because you may be doing an intervention study which causes only one person to overreact at the dose you are testing. If the dose was a bit higher, more people might overreact. If you toss out that subject, whoever does the next study at the next higher dose won't realize that the few outliers they get could be combined with your single outlier to begin to form a dose - response curve for overreaction. This failure has led to disastrous results for several new drugs.

4. Trimming and outright drastically changing some of the numbers so they come out the way you want them to be: The temptation to change a few numbers so they study comes out the way you know it should can be nearly overwhelming. My laboratory recently completed a study in which we found a difference between the control and experimental treatment groups when the first third of the data were gathered. I told the sponsor about the initial findings and they not only sent the rest of the money for the study but tentatively agreed to very substantial funding for subsequent studies if the current study was completed with similar results. Once the rest of the data were gathered, the difference disappeared. It didn't just disappear statistically. There simply were no detectable differences between the groups. It turned out that the technician had not been including all of the subjects who should have been eligible for the study because of a misunderstanding of how to perform the test which determined the entrance criteria. The subjects with relatively mild conditions were systematically excluded from the study. I knew that the groups would have been different if they had been included. The technician went back to college after completing data collection and I was the only person who had the raw data. All I would have had to do to show a difference would have been to change a few of the readings so they fit with what I knew they should have been if the study had been done correctly. Those few strokes of the keyboard would have given me an important paper and lots of funding for further studies during which I could have corrected the mistake and come up with the real results. Instead, I informed the sponsor of the real result and my suspicions concerning what had gone wrong and how to correct it. Of course, I hoped that they would give us sufficient funding to redo the study sufficiently to substantiate (or not) my suspicions. They didn't.

5. Pretending that other work supports yours even though it doesn't or failing to read the other work and then quoting it: The best example of this that I know of was one I came across myself when doing background work for a study on the MMPI (Minnesota Multiphasic Personality Inventory). People who are exaggerating their back pain are supposed to produce a pattern of responses called a \square conversion v. \square . Literally dozens of articles quote the person who first published on this. However, the article that all of these dozens of people quote doesn't even mention it. Instead, the work was reported in the author's thesis - which was not published. Everybody knew who established the concept so a few people must have done a literature search, found his name and figured that must have been the article where he presented his idea and supporting data. They were wrong. Everyone else simply read the early articles and made believe they had read the original. Its too bad that nobody ever went back to the original thesis because it would have saved over twenty years of misunderstanding the concept and, indirectly, damming thousands of low back pain patients to incorrect assessment.

If you want to reference something that you read in another publication and don't feel a need to go back to the original source, simply say something like "Effit 1207 as quoted by Mabit 1861". In this example, nobody would be surprised that you couldn't find Effit's work. However,

if Effit wrote two years ago in a major journal in your nation, it would be strange indeed if you thought the work was worth quoting but you couldn't take the time to go find it to insure that the publication which brought the work to your attention was interpreting it correctly.

What happens if you can't replicate someone's results? Does that mean they faked their results? Certainly not! It does mean that you should look very hard for reasons the results may differ. This has happened to me several times. In one case I finally called the senior investigator of the other study and we worked out the problems together. In another case, I eventually visited the lab but never could find out why our results differed. In the first instance, it turned out that both my group and the other team thought we were dealing with similar populations of healthy young males. Mine turned out to be considerably healthier with far less significant histories of foot problems - and we were measuring feet.

In the second instance, one of my students decided he wanted to investigate an aspect of extrasensory perception (ESP). According to a published article, students taking a math test supposedly did better if a sealed envelope containing a test with the correct answers filled in was attached to the test they were taking than they did if the envelope contained a test with the wrong answers filled in. We followed the experimental design exactly as stated in the article and found both groups to do equivalently. Our subjects were college freshman rather than high-school students but they took a similar test. Our failure to replicate got us into the middle of the usual argument about people who don't believe in ESP (or whatever) not getting results that support it while people who do believe in it (what ever the it is) do find support for it. People doing ESP research are frequently accused of performing pathological science. This is the idea that the investigator is so tied up with finding support for the concept that normal procedures are ignored, studies which can not actually prove the point are repeated ad nauseam, and the design which would prove or disprove the point is never actually run. Failure to replicate should not be mixed with the likelihood of pathological science. If a study is done well and the outcomes are objective, any differences have occurred simply by chance or some variation that has not been detected.

Why would health care professionals doing research in an attempt to find ways to improve care for their patients commit fraud??? Knight (1984 as quoted by Levine 1986) and most other people who comment on this problem identify the most pervasive problem as the academic pressure to publish or perish. It is frequently difficult or impossible to publish negative results and negative results rarely lead to continued funding. Thus, your career can be terminated if you don't get positive results. For students, that means the possibility of losing a fellowship or even not being able to graduate from those programs that insist on positive findings before the final series of studies leading to the dissertation are permitted to begin. We have to remember the very powerful human needs (a) to believe you are correct and (b) to maintain your reputation.

D. Plagiarism: Levine (1986) and many others have made it clear that plagiarism is a very serious offense. Normally nobody cares if an unimportant sentence is lifted from an article you read. This can be done unconsciously. However, it is illegal to copy a substantive amount of the work without quoting the originator. This specifically includes copying from an unpublished report. If more than a few paragraphs are going to be used, or if a figure is going to be used, you need written permission from the copyright holder. People frequently plagiarize themselves unintentionally when they use the same paragraphs or even essentially the same paragraphs in

two different publications. For example, I am frequently asked to write reviews of my work on phantom pain. Several areas haven't changed in years and how many different ways can you say the same thing? If I just copy the best way I came up with saying it, I am plagiarizing myself. So, I quote it and get permission from the original journal to do so. An exception to this rule occurs when the work was entirely created by US Federal employees performing their normal duties. This work can not be copyrighted so can be copied by anyone, including the original author, without permission. Of course, it is polite and good politics to get permission anyway if you ever expect to publish in the original journal again.

Students frequently plagiarize work done by others when they write reports or the introduction and methods sections of theses and research protocols. For example, if you were going to do a study on headache activity and happened to read one of the sample protocols at the end of this book that deals with the topic, you might wish to use the section on tracking headaches. That would be fine if you put the section in quotes and give this book credit for it but it would be plagiarism to simply copy it off with a few minor changes in wording.

Stealing people's ideas is, unfortunately, also done by scientists asked to review manuscripts submitted to journals and grants submitted for funding. This is possibly the worst form of theft of intellectual property and frequently lands the thief in court.

One final note about using other people's material: When government employees produce a document as part of their regular duties, that document can not be copyrighted. This means that, legally, anyone can use the material without the author's knowledge, approval, etc. For example, the glossary of research terms near the end of this book was developed by a series of people at Fort Sam Houston as part of their normal duties. I could have reproduced it here without mentioning that they did the work and I certainly didn't need to get their permission to do so. However, this would have been grossly unethical. So, I did give them credit and did get their permission (in writing).

E. Ethical use of experimental / unapproved devices:

(much of the discussion of classes of devices and their approval process comes from Sherman 1994 and Fuson et al 1997)

In the United States, medical devices are divided into three classes in accordance with Federal law (FDA, Medical Device Amendments 1976 to public law 94-295-90 Stat. 539).

1. Class I devices are exceedingly low risk and require no performance standards or special controls and are exempt from most FDA restrictions.

2. Class II devices are also low risk but require conformance to special controls such as labeling requirements, post-market surveillance, and performance standards. Biofeedback machines are examples of Class II devices.

3. Class III devices are relatively high risk devices or those using a new technology or a new application of existing technology. Thus, a totally harmless device can wind up as Class III if the FDA chooses to apply its regulations very strictly.

If a device is substantially equivalent to a device which was marketed before the 1979 act, many Class II devices can be cleared for sale by submitting a "510-K" application to the FDA.

A 510-k is a notification provision by which a company advises the FDA that the device it wishes to sell does not require a premarket approval application (Fuson et al 1997). If no studies have been done to support the use of a new application of an accepted device, it costs well over a hundred thousand dollars to perform the studies required to extend the device's label to cover that application.

If a potentially risky device is essentially new technology, the sponsor of the study (usually the manufacturer) must submit a request for an Investigational Device Exemption to the FDA. This application lets the FDA know the exact study designs you intend to use to show efficacy and safety. They will require modifications as they see fit. Unfortunately, there is no exception for small/ pilot tests of new devices. A new orthopedic device may require seven years to work its way through the approval process at a cost of \$700,000 to \$1,000,000 (Fuson et al 1997). This has resulted in many new devices being tested and used in Europe until sufficient data is available to make it worth getting them accepted by the FDA.

The following material is from an article I wrote concerning the FDA's ability to regulate the sale and use of biofeedback devices (Sherman 1994). These are less risky than most devices so they restrictions on them are very exemplative of the current regulatory situation. The crucial, bottom line is that (a) you must tell a patient when you are using a device outside of its label and (b) you can not charge a patient for a therapy using a device outside its label during a study in which subjects may not benefit from participation.

The FDA is not likely to enforce every aspect of the law as they interpret it because they are already overwhelmed with work many contacts at the FDA do not feel that biofeedback is highly likely to harm anyone. Thus, it remains a very low priority unless someone does something so far out of line so publicly and blatantly that they can't ignore it. Many of our members feel that these regulations are wrong while others feel that they do not apply to some or all aspects of the use of biofeedback devices. For example, when biofeedback devices are used for non-medical uses, such as education not involving clinical conditions, they probably do not fall under the FDA's regulations.

The sale and use of biofeedback devices is governed by FDA regulations. The most relevant is 21CFR 882.5050. The two main issues covered by the regulation which I feel we should be most concerned with relate to (a) what applications can biofeedback devices be advertised (labeled) as being used for and (b) who can buy and use biofeedback devices.

Labeling of biofeedback devices:

In order to get approval to sell biofeedback devices designed after 28 May, 1976 (the date the law regulating medical devices went into effect), a manufacturer has to demonstrate to the FDA that the device is essentially equivalent to one already approved or grand fathered in by having been marketed prior to 28 May, 1976. This is a pre-market notification performed under section 510K of the above regulation and has come to be referred to as a "510K." Please note that the application includes an "intended use" statement or literature which describes what the device is used for. The "labeling" of a device includes all newsletters from the manufacturer, advertising, written material sent with a device, etc. The only intended use statement or "label" for open sale devices I have been able to find with the help of several people from the FDA is one that states that biofeedback devices are approved for "relaxation" (of the stress control type). Several

manufacturers have had "biofeedback", "muscular rehabilitation", or "treatment of fecal incontinence" accepted as the intended uses. No other claims have been approved and, thus, any other "labeled" claims are illegal. Open sale biofeedback devices can not be legally advertised to "cure" hyperactivity, urinary incontinence, etc. unless the manufacturer has filed the claim with the FDA.

A manufacturer does not get "510K" approval as an entity or for an entire line of products. In other words, because one device manufactured by a company has 510K approval, that does not mean others do. It also does not mean that because a manufacturer was in business selling biofeedback equipment prior to May, 1976, that it has approval to continue selling all equipment. Some particular, individual items manufactured prior to May 1976 may be grandfathered in but "first cousins" introduced afterwards are not. The definitions of equivalent and how much you can change an older device and still have it fit the grand fathering provision are up to the FDA. They can be picky but are subject to pressure from ombudsmen and potential law suits.

If a manufacturer wishes to have a different intended use for a product which either (a) already has 510K approval or (b) for which such approval is being applied for, the manufacturer must supply clinical efficacy studies which can convince an FDA scientific review panel of their validity. If they are convinced, a "hybrid" 510K can be issued. Please note that because biofeedback devices are not used to support or sustain life, are not implants, and do not put the subject at "unusual" risk, the studies do not have to have been pre-approved by the FDA prior to being performed to be considered by the panel. However, they do have to have been approved by a legally constituted Institutional Review Committee (IRB). This is a time consuming and relatively complex action to attempt but is much easier than the following alternative. I feel that we clearly have the data which would support altering the intended use statements for several devices. This is especially true of devices marketed for EMG biofeedback for amelioration of tension headaches.

If a device is not essentially equivalent to an approved or grandfathered device, then "pre-market approval" (PMA), as opposed to the pre-market notification discussed above, has to be applied for under section 515 of the Act. This is much tougher to get. Many biofeedback devices may not be the equivalent of devices marketed prior to 1976 and probably require this kind of approval. Much of our software could potentially fit into this category.

The starting point for approval of new devices is usually getting an "Investigational Device Exemption" (IDE) from pre-market notification and pre-market approval so experimental devices can be supplied to researchers for clinical trials on safety and effectiveness (Section 520(g) of Title 21, FDA 90-4159). These are relatively easy to get but the device's use is very strictly regulated and clinical trials must frequently be pre-approved by the FDA before their data can be used to support an application for efficacy. Specific safeguards for humans who are subjects of investigations, maintenance of sound ethical standards, and procedures to assure development of reliable scientific data are required of EACH user. Studies must be approved by an Institutional Review Committee! Although manufacturers frequently fill out the paperwork for an IDE, the individual investigator is named in and is responsible for it.

There is an exception to the requirement for getting an IDE. If, and only if, a legally constituted IRB (Operating under the guidelines of part 56 of 520(g) and, usually, registered with the US Government) states that the device and study present no significant risk to the subjects, no IDE submission to the FDA is required. However, the definition of "significant risk" includes the concepts of welfare and important treatment rather than being limited to physical damage. The relevant portion of the regulation is: ".. presents a potential for serious risk to the health,

safety, or welfare of subjects, is for a use of substantial importance in diagnosing, curing, mitigating, or treating disease or otherwise preventing impairment of human health ..." This provision should hold even though many biofeedback devices meet other requirements for exemption as most are noninvasive, do not by design or intention introduce energy into a subject (except GSR devices), and are not used as a diagnostic procedure without confirmation by another medically established diagnostic procedure . Thus, I believe that all biofeedback devices fall within the requirement for an IDE.

The IDE and IRB approval can not be used as a cover to sell unapproved devices for supposed "research." I do not believe that any biofeedback device can be sold labeled as being "for research only" (the required wording is "CAUTION - Investigational Device. Limited by US Law to investigational use") and meet the requirements of the law unless each purchaser has an IDE or, if a non-significant risk device, IRB approval prior to sale. Prolongation of the study and commercialization and promotion of an investigational device are strictly prohibited. In addition, "the manufacturer must not charge the investigator more than the cost necessary to recover the costs of manufacturing, research, development, and handling." In other words, no profits!

Sale and use of biofeedback devices:

There are two ways biofeedback devices can be sold. According to FDA representatives, they can be "class two non-prescription devices" only if their single stated ("labeled") use is for relaxation (as in stress/anxiety reduction - not muscular relaxation for control of abnormal muscle tension).

They are "class two prescription devices" if their stated ("labeled") uses are for or include any other purpose. If there is an other stated purpose, they can not be used unless they are "prescribed" by someone competent to do so. This can include licensed clinical psychologists depending on State regulations. A device is classified as "prescription" "if adequate directions could not be prepared for the lay person" (21CFR) to use the device. Thus, logically, they should not be sold to the general public. For example, Biosearch's fecal incontinence pressure biofeedback devices were described in the 510K as being for "biofeedback training of anal sphincters for people with incontinence | sold to licensed physicians and psychologists.

It is clear from the above information that most biofeedback devices logically fall into the "prescription" category. Once the FDA takes a careful look at what we use biofeedback devices for, I fully expect the FDA to change the classifications of all but a very few devices to prescription.

The FDA has the power to stop the sale of improperly advertised devices. It can seize them from manufacturers and practitioners and / or fine the manufacturers if it wishes. As increased attention is drawn to biofeedback, the FDA is very likely to take a closer look. When they do, they could require major changes immediately.

FDA regulation of individual practitioner's use of biofeedback devices:

Numerous members of the AAPB have asked what the ramifications of the above information are for individual therapists using biofeedback devices beyond their legal labels. The FDA definitely has the right to stop any practitioner from using any device it deems dangerous. When a representative of the FDA was asked to comment about this issue, he said that "the FDA does not regulate the practice of medicine" but they do protect patients. I asked him the following questions and received the associated answers:

a. Can licensed practitioners use biofeedback devices for non FDA approved uses? This

constitutes adulteration and is essentially a "new intended use". If approval has not been granted from an institutional review committee (IRC) and, frequently, an investigational device exemption (IDE) has not been granted, the provider is at great risk of being sued by the patient if something goes wrong. Using the device outside of its label is similar to having a patient use a prescription drug for non-approved purposes. The patient should at least be informed that the device is being used outside of its label. The manufacturer should file either a new or amended / hybrid 510K to have a substantiated use added to its label.

b. Can the FDA seize devices from practitioners using them for non-approved uses/therapies and fine them for using them in this way? YES. Commissioner Kessler revisited" this issue for other types of devices in 1994. This is essentially an "illegal use of a device" and the FDA could theoretically intervene. They do intervene if patients are in danger. The representative said that since biofeedback devices are unlikely to cause harm it is unlikely that they will intervene.

It is clear from these responses the most biofeedback practitioners are putting themselves at some risk by using devices beyond their legal labels. It is also clear that the FDA is not likely to intervene unless they feel that patients' safety is at risk. This does not mean that research can not continue or that workshops on specific uses can not be given. It does mean that practitioners need to use their devices as labeled if they wish to remain within the letter of the law.

F. Ethical considerations on incorporation of nonvalidated / innovative diagnostic and therapeutic techniques into regular practice on a recurring basis and billing for them:

This is a complex, highly controversial (not to mention intensely emotional) topic. The central issue concerns the circumstances under which private practitioners can provide and charge for a nonvalidated diagnostic or therapeutic approach on an ongoing basis (as opposed to a single trial) which is permissible by law and within the scope of their licence and expertise and which they believe has a **reasonable** chance of helping the individual patient being considered for it. The crucial question is how to make the legally and ethically supportable decision that the approach has a **reasonable** chance of helping that individual patient.

The following chain of decisions and documents could be interpreted in several ways and can be considered more or less binding on any individual practitioner depending on each State's laws and statements contained in each professional organization's ethics documents. Many private practitioners feel they have far more freedom under their State licences than is indicated here but that concept is continually being challenged in court.

Intrinsic to this discussion is the concept that just because it is legal to perform or charge for an intervention does not indicate that it is ethical to do so. The overlap between innovative therapy and research is quite problematical. It has been discussed by such workers as Levine (1986) and guidelines for providing nonvalidated practices have been issued by a presidential commission (under Public Law 93-438). The commission's decision about nonvalidated practices is as follows:

"Uncertainties may be introduced by the application of novel procedures as, for example, when deviations from common practice in drug administration, or in surgical, medical, or behavioral therapies (*or evaluations - au*) are tried in the course of rendering treatment. These activities may be designated innovative therapy, but they do not constitute

research unless formally structured as a research project. There is concern that innovative (*i.e. nonvalidated - au*) therapies are being applied in an unsupervised way, as part of practice. It is our recommendation that significant innovations in therapy should be incorporated into a research project in order to establish their safety and efficacy while retaining the therapeutic objectives."

Levine (1986) feels that "a practice might be nonvalidated because it is new and has not been tested sufficiently often or sufficiently well to permit a satisfactory prediction of its safety or efficacy in a patient population." or a frequently used practice turns out to never have been appropriately validated. The bottom line is that you are not supposed to keep using nonvalidated practices indefinitely. They are to be tested appropriately and then either dropped or accepted. Many biofeedback and surgical approaches fall into this category and their continued use on a "clinical" basis without adequate testing clearly contravenes generally accepted ethical guidelines and the FDA's regulations, as well as possibly infringes on several laws. This is an area rife with potential for government action and law suits from dissatisfied clients.

The Belmont report (1988) makes the crucial distinction between practice and research as follows: "For the most part, the term "practice" refers to interventions that are designed solely to enhance the well-being of an individual patient or client and that have a reasonable expectation of success. The purpose of medical or behavioral practice is to provide diagnosis, preventive treatment or therapy to particular individuals. By contrast, the term "research" designates an activity designed to test a hypothesis, permit conclusions to be drawn, and thereby to develop or contribute to generalizable knowledge. When a clinician departs in a significant way from standard or accepted practice, the innovation does not, in and of itself, constitute research. The fact that a procedure is "experimental", in the sense of new, untested, or different, does not automatically place it in the category of research. **Radically new procedures of this description should, however, be made the object of formal research at an early stage in order to determine whether they are safe and effective. It is the responsibility of medical practice committees** (*defined elsewhere as professional organizations, State boards, etc. - author*) **to insist that a major innovation be incorporated into a formal research project.**" Thus, virtually all new interventions as diverse as a new surgical procedures to a non-invasive behavioral intervention need to be tested before they are put into general practice.

There is little disagreement about not charging patients who may not benefit from a therapeutic approach who are being formally tested in a placebo controlled study. In fact, it is appropriate to pay subjects when there is a possibility of risk. The controversy surrounds using a treatment which is not generally accepted as efficacious and charging patients for it. The problem is that even if you believe that the treatment is very likely to help that individual patient, you are probably using a device or drug off label and are opening yourself to both professional and legal action. The days when health care providers could do anything they happened to think up as long as it was within the scope of their State licence (practice expertise) simply because they felt it would help an individual patient are disappearing in the United States, Canada, and most European nations. Federal laws supersede any activities permitted under State licensing. If you work in a hospital or other health care organization, you must obey the organization's treatment algorithm for each disorder (standards of care) or explain, in writing, why you deviated

from them. Private practitioners are not as closely scrutinized YET except by insurance carriers who may not wish to support the provider's decision about choice of intervention. Professional organizations and state credentialing bodies rarely intervene unless the provider is clearly doing very odd activities with considerable attendant publicity or doing harm.

When you elect to utilize a nonapproved / innovative approach, you **must** at least warn your patient (in writing) that you are using a device or drug off label. You must tell the patient that the intervention you are proposing is not normally accepted for use for that condition or, possibly, any condition. Obviously, you will discuss the odds of improvement and any known risks with the patient as you would for any accepted intervention. *Providers tend to charge for treatments within their scope of practice and expertise which they feel are reasonably likely to help an individual patient for which they have sufficient evidence to satisfy themselves that the attempt is warranted in balance with the risks of performing and not performing the intervention.* This decision is usually ethically and legally defensible if the average person would make the same decision with the same information using principles such as those espoused by Evidence Based Medicine (discussed elsewhere in this book).

The minimal requirement for continuing to use an innovative / nonapproved practice appears to be to collect and report systematic data about its effectiveness and, of course, to inform your patients that you are doing so. These minimally intrusive □research projects□ are intended to avoid compromising optimal clinical care in any way. They simply require the application of excellent clinical practice to a group of similar patients. In such a design, an appropriate diagnostic procedure must be performed, an adequate baseline taken, the intervention must be well defined, and an adequately long follow-up must be performed so that normal variations in the disorder can be detected. **It should be specifically noted that there is no requirement to perform double-blind, placebo controlled, cross-over studies in order to demonstrate the efficacy of a procedure. While such designs are occasionally appropriate for some short acting drugs, they are rarely appropriate for surgical and behavioral interventions.** These types of studies are discussed in detail in the experimental design portion of the book.

Thus, you can not consistently expose people to the light of the full moon to cure their broken legs for very long without risking severe legal and professional criticism unless you (a) can present some reasonable rationale for the attempt and (b) incorporate the innovation / nonapproved practice into a clinical study.

Given the above constraints, the question arises as to how the practitioner should decide (a) whether a novel intervention is sufficiently well demonstrated to warrant charging for it and then (b) when it is sufficiently well accepted so it is no longer considered innovative. There is no simple answer because the FDA, professional organizations, insurance carriers, and health care organizations all use different, conflicting, idiosyncratic criteria for □accepting□ an intervention as efficacious. Many organizations such as the American Psychological Association have adopted requirements such as the following for determining that a treatment has been shown to be efficacious:

- a. Two studies with appropriate design and controls (group design or a series of single case design studies).
- b. Studies conducted by different researchers.
- c. Studies demonstrate clinical efficacy. The new treatment must be shown to be efficacious in comparison with medication, a placebo, or another treatment. The treatment must

be shown to be equally effective to an established treatment for the same problem.

- d. Waiting list controls are not sufficient for demonstrating efficacy.
- e. The diagnostic criteria must be clearly specified.

To reiterate, there is no requirement that these studies contain double-blind, placebo controlled, crossover designs.

There is no particular relationship between an intervention being accepted and being substantiated as effective. In fact, the matter of general acceptance of an innovative / nonapproved approach by the medical administrative and clinical communities is, unfortunately, not well integrated with demonstration of efficacy. Rather, it is a political process which must be discussed because ethics tend to be forgotten in the struggle for the institutional reimbursement which accompanies general acceptance of an approach. Incredibly enough, in the United States the following seems to be the most common way a novel approach becomes reimbursable:

- a. Get lots of positive publicity through manufacturers' representatives, newspapers, TV, endorsements by famous (non-medical) people, etc.
 - b. Have the treatment represented by a well known, successful, senior provider of the type highly esteemed by the organization you are attempting to influence. This is part of establishing the innovation's credibility.
 - c. Know and play the organization's political game exactly correctly with infinite patience.
 - d. Induce a large group of the organization's clients clamor for the treatment.
 - e. Provide solid evidence that using the treatment will save the organization money in a very short time.
 - f. Provide an acceptable level of evidence that the treatment meets efficacy criteria so the organization can support its decision to pay for it.
- Please note that the ethics involved in the acceptance process are marginal at best.

The conclusion to this convoluted mess is relatively straightforward:

1. It is appropriate for clinicians to use every ethical, legal means within the scope of their license and experience to provide diagnostic and therapeutic interventions for individual patients which have been demonstrated to have a reasonable chance of being efficacious for that patient. Practitioners can charge for these interventions.
2. There are accepted criteria for judging an innovative therapy as having been demonstrated to be efficacious. Practitioners must base their decisions about likely efficacy for their individual patient upon these criteria rather than personal feelings.
3. Practitioners wishing to use innovative interventions more than a few times must test that intervention for efficacy according to accepted methodology.
4. It is unethical for practitioners to build nonvalidated interventions/techniques into their practices on a regular basis or to charge for them when so included.
5. It is incumbent upon clinicians to know enough about experimental design and analysis to know when techniques have met efficacy criteria.
6. For the third time, it is emphasized that there is no requirement to perform double-blind, placebo controlled, cross-over studies in order to demonstrate the efficacy of a procedure. While such designs are occasionally appropriate for some short acting drugs, they are rarely appropriate for surgical and behavioral interventions.

G. The complaint investigation process:

Conflicts between any of the involved parties occasionally arise during the conduct of a study. For example, a technician may feel that the data are being faked, that subjects are being harmed, etc. A member of the investigator's department may feel that the idea was stolen or that the study is being performed differently than approved by the protocol. The question is what these people can do to **effectively** voice their complaints without risking their livelihoods, positions, etc. A large body of legal decisions has built up over the years which continuously tests the laws and ethical guidelines governing this topic. The accepted practice was codified in a book (President's 1982) based on an agreement reached by the American Association for the Advancement of Science, Medicine in the Public Interest, and the President's Commission for the study of ethical problems in medicine and biomedical and behavioral research. The basic principals which govern investigation of misconduct are (President's 1982):

1. Institutions must have a specific office designated to receive and investigate complaints. At small institutions this tends to be the office of the director of research (or whatever that position is called). Large institutions usually have a quality control office or an inspector general which would handle reception of the initial complaint.
2. Mechanisms for assuring a prompt investigation must be published and in place. Normally, the director of research or the head of the office which receives such complaints will either take the complaint personally or have someone regularly assigned to receive them.
3. An impartial adjudicator must be assigned. The responsible official must assign an investigating officer who is uninvolved with the situation but knowledgeable in the field within days of receiving the complaint.
4. The investigating officer must conduct a through investigation as rapidly as possible. Full opportunity for the complaining parties and the accused to explain their positions, present evidence, call witnesses, and so on must be provided. Every attempt must be made to keep the identity of the complainant anonymous. This can get complex in small institutions because it may be obvious who complained.
5. Protection from reprisals for the good-faith complainant and for witnesses must be provided. If the investigation determines that there was no problem and that the good-faith complaint was the result of a misunderstanding or etc., absolutely no action may be taken against the complainant. It must be emphasized that if the investigation determines that the complainant was not acting in good faith, then evidence of a crime has been uncovered and the institution is very likely to seek legal redress. If the investigation determines that there really is a problem, regardless of whether it is a minor lapse of ethics, a legal issue, or whatever, it must be dealt with and corrected rapidly. It need not be dealt with publicly unless a law has been broken. Institutions can take academic and legal steps against perpetrators. Academic steps are taken within the institutions regulations for handling such problems. Many institutions have a special faculty committee for adjudicating such situations with their decision appealable to the head of the institution.

Institutional Review Boards (IRBs) should not be expected to perform monitoring,

investigative, or adjudicative functions. Applicable regulations should be clarified as to what is intended (and not intended) by the charge to IRBs to perform "continuing review" and to report serious and continuing noncompliance. Reasons given in support of this recommendation include the fact that IRBs do not have the time, the resources (staff, money), or the expertise to perform such functions. In addition, adoption of the monitoring role would conflict with the primary role of IRBs: to educate and advise research scientists and to resolve problems in a constructive way. Finally, many, if not most, institutions already have appropriate quality assurance mechanisms in place.

IRBs should be kept informed of all allegations of misconduct in research with human subjects and of investigations, as well as findings, relating to such allegations. The IRB might be consulted as to the seriousness of the misconduct found to have occurred.

Institutional administrators, principal investigators, and research personnel must be made aware of their responsibilities to the scientific community and to federal agencies.

Serious misconduct must be reported to the cognizant federal agency after a formal determination has been made by the institution. Administrators and scientists should understand that they have a legal obligation to do so. In fact, to knowingly provide false information to the federal government is a felony. If an institution makes a formal finding that false information has been submitted in a grant application, annual report, or data submitted to a regulatory agency, the institution may incur criminal liability if officials fail to report such a finding.

Academic misconduct in research involving federal funding or FDA-regulated test articles must be reported to the appropriate Federal agency. Levine (1986) states that "the FDA conducts inspections not only for cause but also routinely. In my judgment, their data warrant their conclusion: "As long as the problem continues to exist, the FDA should continue or even expand its monitoring program, while ensuring its rigor." This judgment is not inconsistent with my view that all agents involved in safeguarding the rights and welfare of human research subjects should presume that all other agents are trustworthy and that they are performing competently until substantial contrary evidence is brought forward. Routine audits should be conducted as if investigators are to be trusted. For cause inspections should be conducted differently because they are based upon the availability of substantial contrary evidence."

H. Compensation for injury while participating in a research study:

The laws of chance dictate that eventually a subject will somehow get hurt or think her or she suffered harm or psychological distress while participating in even the most minimal risk study. Someone is going to stick the point of a pencil in their finger while filling out a survey, experience thigh pain while being exposed to a placebo machine (as one of my subjects did), get shocked by a supposedly safe standard device being used to record a physiological parameter, become mentally unstable while performing a relaxation exercise, or have any other event occur that anyone can imagine. The investigator and the institution have to be ready to accept pecuniary liability. If your study is being performed for an organization (such as a drug or equipment company), the sponsor must agree, in writing, to cover the medical costs associated with injury during the study.

The basic legal stance is that "human subjects who suffer physical, psychological, or social injury in the course of research ... should be compensated if (1) the injury is proximately

caused by such research and (2) the injury on balance exceeds that reasonably associated with such illness from which the subject may be suffering, as well as with treatment usually associated with such illness at the time the subject began participation in the research.

For research activities involving more than minimal risk, DHHS and Food and Drug Administration (FDA) regulations both require a statement on the availability of medical therapy as well as compensation for disability.

If you or your institution can not or do not intend to provide compensation, you must say so in the protocol's consent form. If there are no reasonable risks to participation, the consent form must say so. If there is some reasonable possibility of risk, the consent form must give a reasonable estimate of the likelihood of harm, define the harm, and estimate the likely extent of the harm.

I. Ethical issues of using non-human animals in research: Like it or not, under our current laws non-human animals are property - literally "things" with no legal status or "rights" as individual beings. With few exceptions, they can't communicate with us to tell us whether they want to participate in a study. There are no known non-human animals who understand our communications well enough to make an informed consent about whether they should participate in a study. Thus, if an animal is going to be used in a study, humans have to make the decision about whether that animal should participate. We are guided by ethics which have grown up around laws to prevent cruelty to animals and many other considerations.

1. Justification for use of non-human animals in research: Animals are currently a vital, if involuntary, partner in medical progress. Depending on the source you read, between ten and sixty percent of the people who are reading this would have died before adulthood without animal research and a large minority of the survivors would be so debilitated that it is unlikely that they would be learning this material.

Sometimes there just isn't any other way to get clinically vital information so we need to test animals. At this stage in our knowledge, cadavers, computer models and dolls are rarely able to replace a living animal because they don't heal and they do not have the complex systems interactions that living creatures do. If any appropriate animal model can be found which mimics the human response to the situation, such as a goat's knee used as a model for a human knee, the decision has to be made whether to use a goat or a person to get information that is (a) truly necessary to get and (b) can't be gotten from any other source than a living creature.

Many people advocate using prisoners - especially those on death row instead of animals as there is no question that they would be a good model for human physiology and anatomy. This question can be debated endlessly but the fact is that we don't have enough death row inmates to replace all the animals used in crucially important surgical studies likely to make major changes in our lives - let alone those used in medical studies.

Many of the surgical and medical studies which used animals have benefited humans as much as humans. This is a crucial factor to keep in mind when considering whether animals

should be used as subjects.

2. Sensitivity about using animals

a. The ethics of using pound vs. pure bred animals: Do animal researchers really steal people's pet mice to torture? How about their dogs and cats? These days are probably gone in the United States for the foreseeable future. It turns out that dogs and cats are poor models for most surgical studies and that misleading information has been gathered for years by using them. They are still very valuable in many other types of studies. If given the option, most researchers would prefer to use animals bred for research because the history of disease is well known and the animal is likely to be very similar to the other animals in the study - thus reducing variability in response.

Several million dogs and cats are "euthanized" in animal "shelter" gas chambers or with lethal injections every year. These pound animals could be available for use in non-survival (do not wake up) studies. The question is whether these animals who are about to be "terminated" should be permitted to be participants in studies where they will be fully anesthetized and will never wake up after the study. They will not feel any more than they would if they were gassed. They might even experience less terror as they would probably be anesthetized with a shot rather than dying in a mass death chamber. Many people believe that they should be used for research because they would die anyway so they may as well help humanity while they are doing it. This attitude puts the blame for the problem squarely on the heads of the millions of Americans who abandon their animals every year or permit their animals to breed freely in the wild. Many animals are abandoned on the roadside when the owners move or when the animals become sick or elderly. The number of people who lose and then can't find their animals is tiny compared to the above scenarios. Thus, the animals killed in shelters are rarely anyone's beloved pets. They are feral or abandoned creatures.

b. The "situation" with relatively intelligent animals such as dogs, cats, monkeys, and dolphins: We all know that plenty of animals are smarter and nicer than we are in their own ways (and occasionally in ours as well). How can we condone the use of an intelligent creature with emotions which clearly mirror our own for studies which could cause pain, terror, and death? The pragmatic answer that virtually all animal use committees and regulations use at this time is that the least "advanced" type of animal which can fill the need must be used. Thus, if a hamster can do the job as well as a dog, you need to use the hamster. There are also many restrictions on the living conditions for relatively intelligent animals. For example, monkeys need a minimum amount of social contact and living space. The amount of pain, terror, etc. they can be exposed to is strictly limited.

c. Pain and multiple surgeries: There are now strict rules against torturing animals. Animals must be given at least the same amount of anesthetic and analgesic which would be used on a human patient under similar circumstances. If an animal has to experience pain for some reason essential to the study (such as studies on pain relief methods), the intensity and duration of the pain must be minimized and the very minimal number of animals must be used. It takes quite a bit to get one of these studies through an animal use committee. The investigators have to prove beyond any likely doubt that the pain is really necessary and that the study is likely to actually have significant benefits to both humans and non-human animals.

Animals are usually not permitted to be used for multiple surgical procedures from which they will awaken because of the stress it puts on them. Investigators who wish to use animals in this way must justify their request as well as for painful studies.

3. Factors entering the decision to use an animal model:

- a. Is there a way to find out what you want to know in vitro, with computer models, or in humans (without harming them)?
- b. Has the work already been done? - Strategies for literature searches.
- c. Is there an adequate animal model for what you want to investigate?
- d. Physical and mental cost **to** (not of) the animal vs. potential benefit to society.
- e. Pain, acute and chronic, the animal will have to undergo. Pain must be minimized to the fullest extent possible.
- f. Whether the animal has to undergo multiple survival surgeries. These studies are rarely approved anymore.
- g. Whether the animal will survive after the study.
- h. The ratio between the number of animals required and the likely benefit of the study to humans and non-human animals.

4. Design and performance of ethical studies using non-human animals:

To perform a study with non-human animals in the current ethical environment, great care has to be taken to insure that the correct animal model is used, that the minimal number of animals required to answer the question participate, and that the design is optimized so the odds of getting a meaningful answer are as high as possible. Everyone is urged to follow the four “Rs” of animal research: Replacement (use the correct model), Refinement (use the optimal design to get the best answer), Reduction (use the least possible animals), and Responsibility (everyone associated with the study - from animal handlers through the head of the institution - continuously demonstrates concern for the health, comfort, welfare, and well-being of the animals in their charge.)

a. The study must have significant value to humanity and the animals: Above all else, the question has to be worth answering. Several years ago my institution was approached by a group of plastic surgeons from both our institution and a local medical school who wanted to perform a new skin flap sectioning procedure with pigs. Upon extensive questioning it finally turned out that the procedure would be of no value for anything other than making face-lifts around the eyes a bit nicer. The knowledge gained about skin flaps and underlying fatty tissue would not be translatable to any other procedure such as one which might help trauma or burn victims recover with more sightly skin. Even as recently as a decade ago, this study might have been approved. As it was, we turned them down because it simply wasn't worth the pig's lives. It later turned out that the reason they approached our institution was that the medical school had already turned them down.

b. Use the minimum number of animals: Just as with human studies, you must not use more subjects than you need to get the answer. If you have any indication of expected variability in the response, the magnitude of the response and the difference / change which would be clinically important, you can perform a power analysis to give you an indication of

how many animals you will need to have a reasonable chance of achieving a significant difference. These techniques are discussed later in the book. If you have no idea what to expect, you need to either conduct a pilot study with a few animals to find out what the variability and average response will be or conduct the study in such a way that only a few animals start at a time so it can be stopped when enough animals have participated to give you the answer you need.

The days are, thankfully, gone when you could do one mass exposure to find out the dose of some chemical which would kill off half the animals exposed to it while simultaneously finding the dose that would kill all those exposed and the dose which would not kill any of them. This type of study was called an LD/50 (50% live, 50% die). I have stark memories of studies in which a hundred unconscious rats would be spread out in neat lines on a huge table. Each group of five had received one of a gradually progressing dose of the drug being tested with the minimum dose being enough to at least knock out the animals. Of course, most of them received the wrong dose so died for no reason. Occasionally the whole group received too low a dose and the study had to be redone with another hundred victims.

The appropriate design for this test was well known but ignored because it took more time. The first step is to take a few animals and do a dose ranging study. If an animal gets really sick or dies, you know the range to use for the rest - but you gradually increase the dosage until you get to the point you need. This may take only fifteen or twenty animals but takes days or a week instead of hours.

c. The correct model must be used: If you truly can not find a non-animal model, then use the optimal animal to give the answer you need. When choosing this model, you must choose the animal lowest on the scale of self-awareness possible. Dogs and cats used to be the favorite animals to use in studies because of their easy availability. Countless orthopedic studies were performed using dogs before it was learned that their bones and joints are poor models for many human problems. It turned out that goats and pigs were the optimal models because of weight bearing patterns, joint structure, etc. I wish I could say that the switch to goats and pigs was made because they were better models but that wouldn't be the truth. The switch was made because of political pressure not to use dogs. As surgeons got more experience with goats and pigs, the models turned out to give better answers so the whine level gradually decreased.

You are not expected to be able to figure out which model is best for your study on your own or from reading the old literature. Most animal facilities have a specially trained veterinarian either on contract or in place who has an extra four year residency in laboratory animal care. That person is trained to help you sort out the literature and find the optimal model. Literature searches of data bases which specialize in identifying animal models, such as the Department of Agriculture's, are also critical.

Your choice must be firmly justified with facts. You can't say that you need to use a goat because two recent studies in your field used them. You need to explicitly show why that goat joint (or whatever) is better than a rat's, etc. If the rat's joint is too small to see or bears weight differently, has a different blood supply, etc. that is acceptable - but you have to say it.

d. The study design is crucial! Too many studies are so poorly designed that they waste animals or can't really answer the question they are asking. Many studies are far more efficient when done in stages because the design can be modified as you go. For example, a study determining which coating would result in optimal ingrowth of bone into a metal rod

required drilling 3 mm diameter holes into goat's long bones so coated rods could be inserted into the bone. The idea was to get as many rods into a goat as possible to save on the number of goats. The surgeons knew that if the holes were too close together, the bone would break so they limited themselves to five holes per bone and only did one bone per goat to insure that the goats could still walk. What they didn't expect was for the holes to act like perforations on a piece of paper and join each other - and that is exactly what happened. They had designed their study to be done in stages so they could get experience with the first few goats. Several of them wound up with broken legs. Because the study was staged, only a few goats had to be repaired with plates rather than an entire 48 goat study being destroyed.

The study design has to be very detailed and needs to include exactly how you intend to treat pain, exactly how you will perform the procedures, where you are going to get the animals and how you will house them - exact specifications of food, water, bedding, etc.

Of great importance, you also need to demonstrate that you know how to do the proposed procedures or that you are working under the direct supervision of someone who does. This directly translates into restricting non-surgeons from doing surgery on animals unless that non-surgeon is appropriately trained and supervised. This simple rule has prevented many of the worst abuses animals had to suffer due to incompetent bungling.

The study design section of this book will guide you toward thinking out an optimal design and the required components of a typical study.

e. Ongoing responsibility for your charges: The animals in your study can not protect themselves. The study must be designed to minimize their suffering. The rule is that if a person would suffer under the circumstances, so will an animal. You must provide at least as much, if not more, alleviation for that suffering than you would provide for a human. Animals must not experience pain unless the aim of the study is alleviation of pain - and it must be minimized even in that case. The study must be designed so a sick or suffering animal can be euthanized promptly if it can not be treated successfully.

Very strict laws have been established by Congress (The Animal Welfare Act - 7 USC 2131-2156 etc.) and promulgated by numerous government agencies which govern use of animals in research by anyone in the United States regardless of the nature of the facility they work in. The act mandates the establishment of a Laboratory Animal Care and Use Committee (LACUC) at every facility using animals and gives that committee the authority to govern use of animals at that facility - including stopping a study dead in its tracks if they do not like something happening in the study. Nobody in the facility can overrule the committee's decision to stop a protocol or to not approve it in the first place. Everyone at the facility has the right to report problems to the facility without retribution. The veterinarian assigned to overlook the animal's care for that study has, through the LACUC, the power to order a study stopped instantly or to order any particular animal removed from the study for care or euthanization.

Laws govern such details as how large a cage has to be for each animal of any given species, crowding, transport, social conditions, health and comfort levels, etc. For example, primates need very large cages, have to have compatible company for some percent of their waking hours, and need a minimum amount of positive intellectual stimulation. If our laws for human conditions were anywhere near as strict and well enforced as those for research animals, we would have pretty empty subway trains and a lot less slums filled with sick people.

A crucial last question is what happens to the animals at the end of the study. Many

times, there is no choice but to euthanize (a nice word for "kill painlessly/stresslessly") This is the kicker that stops many people from doing animal research. If you do a procedure which would probably cause the animal suffering and there is no need for the animal to awaken after the procedure, the animal is usually allowed to die while still under anesthesia so it never suffers. However, such non-survival procedures are rare. Usually the animal could live on indefinitely after the study over. In the example about the goat's having rods put in one leg each, after healing, the bone had to be removed so the resistance of the rods to push-out (and bone shattering) could be tested. The goats could have been allowed to recover from the surgery but what do you do with 48 three legged goats? Goats' hooves and legs breakdown if they have to put too much weight on any leg so three legged goats live in apparent agony for a while and finally can't walk any more. Thus, the decision was made to euthanize the goats. As you read the various studies in this book, try to draw your own conclusions about whether the participants should be permitted to survive participation.

J. The Nuremberg Code

Reprinted from Trials of War Criminals before Council Law No. 10, Vol. 2 (Washington, D.C. US. Government Printing Office, 1949), pp. 181-182.

Permissible Medical Experiments

The great weight of evidence before us is to the effect that certain types of medical experiments on human beings, when kept within reasonable well-defined bounds, conform to the ethics of the medical profession generally. The protagonists of the practice of human experimentation justify their views on the basis that such experiments yield results for the good of society that are unprocurable by other methods or means of study. All agree, however, that certain basic principles must be observed in order to satisfy moral, ethical and legal concepts:

1. The voluntary consent of the human subject is absolutely essential. This means that the person involved should have legal capacity to give consent; should be so situated as to be able to exercise free power of choice, without the intervention of any element of force, fraud, deceit, duress, overreaching, or other ulterior form of constraint or coercion; and should have sufficient knowledge and comprehension of the elements of the subject matter involved as to enable him to make an understanding and enlightened decision. This latter element requires that before the acceptance of an affirmative decision by the experimental subject there should be made known to him the nature, duration, and purpose of the experiment; the method and means by which it is to be conducted; all inconveniences and hazards reasonably to be expected; and the effects upon his health or person which may possibly come from his participation in the experiment. The duty and responsibility for ascertaining the quality of the consent rests upon each individual who initiates, directs or engages in the experiment. It is a personal duty and responsibility which may not be delegated to another with impunity.

2. The experiment should be such as to yield fruitful results for the good of society, unprocurable by other methods or means of study, and not random and unnecessary in nature.

3. *The experiment should be designed and based on the results of animal experimentation and a knowledge of the natural history of the disease or other problem under study that the anticipated results will justify the performance of the experiment.*
4. *The experiment should be so conducted as to avoid all unnecessary physical and mental suffering and injury.*
5. *No experiment should be conducted where there is an a priori reason to believe that death or disabling injury will occur except, perhaps, in those experiments where the experimental physicians also serve as subjects.*
6. *The degree of risk to be taken should never exceed that determined by the humanitarian importance of the problem to be solved by the experiment.*
7. *Proper preparations should be made and adequate facilities provided to protect the experimental subject against even remote possibilities of injury, disability, or death.*
8. *The experiment should be conducted only by scientifically qualified persons. The highest degree of skill and care should be required through all stages of the experiment of those who conduct or engage in the experiment.*
9. *During the course of the experiment the human subject should be at liberty to bring the experiment to an end if he has reached the physical or mental state where continuation of the experiment seems to him to be impossible.*
10. *During the course of the experiment the scientist in charge must be prepared to terminate the experiment at any stage, if he has probable cause to believe, in the exercise of the good faith, superior skill and careful judgment required of him that a continuation of the experiment is likely to result in injury, disability, or death to the experimental subject.*

K. World Medical Association Declaration of Helsinki - 1964
(successor to the Nuremberg Code)

Recommendations guiding biomedical research involving human subjects. Adopted by the 18th World Medical Assembly, Helsinki, Finland, June 1964, and amended by the 29th World Medical Assembly, Tokyo, Japan, October 1975; the 35th World Medical Assembly, Venice, Italy, October 1983; and the 41st World Medical Assembly, Hong Kong, September 1989.

It is the mission of the physician (note: this is the broad definition - our □health care provider□ concept) to safeguard the health of the people. His or her knowledge and conscience are dedicated to the fulfillment of this mission.

The Declaration of Geneva of the World Medical Association binds the physician with the words, "The Health of my patient will be my first consideration," and the International Code of Medical

Ethics declares that, "A physician shall act only in the patient's interest when providing medical care which might have the effect of weakening the physical and mental condition of the patient."

The purpose of biomedical research involving human subjects must be to improve diagnostic, therapeutic and prophylactic procedures and the understanding of the aetiology and pathogenesis of disease.

In current medical practice most diagnostic, therapeutic or prophylactic procedures involve hazards. This applies especially to biomedical research.

Medical progress is based on research which ultimately must rest in part on experimentation involving human subjects. In the field of biomedical research a fundamental distinction must be recognized between medical research in which the aim is essentially diagnostic or therapeutic for a patient, and medical research, the essential object of which is purely scientific and without implying direct diagnostic or therapeutic value to the person subjected to the research.

Special caution must be exercised in the conduct of research which may affect the environment, and the welfare of animals used for research must be respected. Because it is essential that the results of laboratory experiments be applied to human beings to further scientific knowledge and to help suffering humanity, the World Medical Association has prepared the following recommendations as a guide to every physician in biomedical research involving human subjects. They should be kept under review in the future. It must be stressed that the standards as drafted are only a guide to physicians all over the world. Physicians are not relieved from criminal, civil and ethical responsibilities under the laws of their own countries.

I. Basic Principles

1. Biomedical research involving human subjects must conform to generally accepted scientific principles and should be based on adequately performed laboratory and animal experimentation and on a thorough knowledge of the scientific literature.

2. The design and performance of each experimental procedure involving human subjects should be clearly formulated in an experimental protocol which should be transmitted for consideration, comment and guidance to a specially appointed committee independent of the investigator and the sponsor provided that this independent committee is in conformity with the laws and regulations of the country in which the research experiment is performed.

*3. Biomedical research involving human subjects should be conducted only by scientifically qualified persons and under the supervision of a clinically competent medical person. **The responsibility for the human subject must always rest with a medically qualified person and never rest on the subject of the research, even though the subject has given his or her consent.***

*4. **Biomedical research involving human subjects cannot legitimately be carried out unless the importance of the objective is in proportion to the inherent risk to the subject.***

5. Every biomedical research project involving human subjects should be preceded by careful

*assessment of predictable risks in comparison with foreseeable benefits to the subject or to others. **Concern for the interests of the subject must always prevail over the interests of science and society.***

6. The right of the research subject to safeguard his or her integrity must always be respected. Every precaution should be taken to respect the privacy of the subject and to minimize the impact of the study on the subject's physical and mental integrity and on the personality of the subject.

7. Physicians should abstain from engaging in research projects involving human subjects unless they are satisfied that the hazards involved are believed to be predictable. Physicians should cease any investigation if the hazards are found to outweigh the potential benefits.

*8. **In publication of the results of his or her research, the physician is obliged to preserve the accuracy of the results.** Reports of experimentation not in accordance with the principles laid down in this Declaration should not be accepted for publication.*

9. In any research on human beings, each potential subject must be adequately informed of the aims, methods, anticipated benefits and potential hazards of the study and the discomfort it may entail. He or she should be informed that he or she is at liberty to abstain from participation in the study and that he or she is free to withdraw his or her consent to participation at any time. The physician should then obtain the subject's freely-given informed consent, preferably in writing.

*10. **When obtaining informed consent for the research project the physician should be particularly cautious if the subject is in a dependent relationship to him or her or may consent under duress.** In that case the informed consent should be obtained by a physician who is not engaged in the investigation and who is completely independent of this official relationship.*

11. In case of legal incompetence, informed consent should be obtained from the legal guardian in accordance with national legislation. Where physical or mental incapacity makes it impossible to obtain informed consent, or when the subject is a minor, permission from the responsible relative replaces that of the subject in accordance with national legislation.

Whenever the minor child is in fact able to give a consent, the minor's consent must be obtained in addition to the consent of the minor's legal guardian.

12. The research protocol should always contain a statement of the ethical considerations involved and should indicate that the principles enunciated in the present Declaration are complied with.

II. Medical Research Combined with Professional Care (Clinical Research)

1. In the treatment of the sick person, the physician must be free to use a new diagnostic and therapeutic measure, if in his or her judgment it offers hope of saving life, reestablishing health or alleviating suffering.

2. *The potential benefits, hazards and discomfort of a new method should be weighed against the advantages of the best current diagnostic and therapeutic methods.*

3. *In any medical study, every patient, including those of a control group, if any, should be assured of the best proven diagnostic and therapeutic method.*

4. *The refusal of the patient to participate in a study must never interfere with the physician-patient relationship.*

5. *If the physician considers it essential not to obtain informed consent, the specific reasons for this proposal should be stated in the experimental protocol for transmission to the independent committee.*

6. *The physician can combine medical research with professional care, the objective being the acquisition of new medical knowledge, only to the extent that medical research is justified by its potential diagnostic or therapeutic value for the patient.*

III. Non-Therapeutic Biomedical Research Involving Human Subjects (Non-Clinical Biomedical Research)

1. *In the purely scientific application of medical research carried out on a human being, it is the duty of the physician to remain the protector of the life and health of that person on whom biomedical research is being carried out.*

2. *The subjects should be volunteers - either healthy persons or patients for whom the experimental design is not related to the patient's illness.*

3. *The investigator or the investigating team should discontinue the research if in his/her or their judgment it may, if continued, be harmful to the individual.*

4. *In research on man, the interest of science and society should never take precedence over considerations related to the well-being of the subject.*

Chapter 7

The research protocol approval process

A. Overview: All organized gathering of information using human subjects which is conducted formally or informally by anyone associated with the institution (including, but not limited to, all paid employees, members, volunteers, and students) must be approved by the institution's Human Use Committee (HUC). Requests for approval to perform research are put in the structured format of a formal research protocol. A typical format for the protocol forms section G of this book and samples of acceptable and poor protocols are in section H. Every academic and granting institution has different, frequently changing, formats for their protocols. Their format and instruction should be available on disk or from a web site. You should not find yourself trying to hand type dozens of pre-printed forms. You will usually be asked to hand in a disk with your protocol on it to ease the review process and save a forest or two.

The review and approval process is intended to optimize the chances of doable, high quality research being performed at the institution. The process is intended to be supportive and friendly rather than rigidly obstructive. Research in the United States operates under strict legal and ethical guidelines. Human use committees must insure that all research conducted within their purviews meets these guidelines. The reviewers and committee members are charged with making sure that the protocol is doable within the investigators' capabilities and resources and likely to produce results worthy of the investigators and subjects' time. Efforts at meeting these goals is likely to result in considerable discussion both during the review process and at the actual HUC meeting. Thus, investigators need to recognize the difference between harassment and actual requirements. Of course, HUC members may not vote on any protocol in which they are involved.

B. Definition and authority: The Human Use Committee (HUC) is a group formally designated by the institution to review, approve the initiation of, and to conduct continuing reviews of research involving human subjects in accordance with state and federal regulations. The HUC has the authority under Federal law and International Treaties to approve, require modifications in, or disapprove any research within its purview. The head or governing body of the institution may disapprove any research project approved by the HUC but can not approve a project disapproved by the HUC.

The authority and functions of human use committees in the United States are largely governed by state and federal laws, federal statutes, and international treaties. These regulations apply differently to private organizations than they do to federal and state supported organizations. Any organization doing research in the United States or which is based in the United States and doing research in other nations is covered by these regulations. These regulations are summarized in the code of federal regulations Title 45 part 46 - protection of

human subjects. The regulations are monitored by the Department of Health and Human Services, the National Institutes of Health, the Food and Drug Administration, and the Office for Protection from Research Risks. Specifications and authority for Institutional Review Boards (Human Use Committees) are delineated in the Public Health Service Act as amended by the Health Research Extension Act, Public Law 99-158. Washington State's laws parallel and sometimes extend those of the Federal Government.

C. Purpose of Human Use Committees : The purpose of the HUC is to assure that in any research performed in relation to the institution:

1. Risks to subjects are minimized (a) by using procedures which are consistent with sound research design and which do not unnecessarily expose subjects to risk and (b) whenever appropriate, by incorporating procedures already being performed on the subjects for diagnostic or treatment purposes.

2. Risks to subjects are reasonable in relation to anticipated benefits, if any, to the subjects, and the importance of the knowledge that may be expected to result.

3. The experimental design and methodology as scientifically sound to optimize the likelihood of producing a valuable result.

4. The protocol meets all applicable ethical guidelines for research and use of human subjects.

5. Selection of the subjects is equitable.

6. Informed consent will be sought from each prospective subject or the subject's legally authorized representative and will be documented in accordance with and to the extent required by applicable laws.

7. When appropriate, the research plan makes adequate provision for monitoring the data collected to ensure the safety of subjects.

8. There are adequate provisions to protect the privacy of subjects and to maintain the confidentiality of data.

9. Appropriate additional safeguards have been included in the study to protect the rights and welfare of subjects who are members of a particularly vulnerable group.

D. Membership of the HUC: The HUC is frequently chaired by the institution's Director of Research. The HUC must have at least five members with various professional and technical backgrounds to promote a thorough and adequate review of research activities commonly conducted by the institution. The HUC must have a diversity of membership (considering occupation, race, gender, and cultural backgrounds and sensitivities to community attitudes in relationship to the research to be performed). Experience and expertise, as well as diversity,

serve to promote respect for the HUC's advice and counsel in safeguarding the rights and welfare of human subjects. Members must possess the professional competence necessary to review specific research activities. Further, members must be able to ascertain (1) acceptability of proposed research in terms of institutional commitments and regulations, (2) applicable laws, and (3) standards of professional conduct and practice.

The HUC must have at least one member in each of these categories: (1) whose primary concerns are primarily scientific (e.g. a professional scientist), (2) whose primary concerns are primarily non-scientific (e.g. an administrator), (3) who is a member of the community representative of the types of subjects likely to participate in the institute's studies (This person and the person's immediate family may not be affiliated with the institution.), and (4) who is a clinician with appropriate expertise in the types of studies performed by the institution.

E. The review process:

1. The review process is intended to strengthen protocols to optimize their chances of producing valuable results. The reviewers are encouraged to take positions as friendly supporters rather than detractors. It is not the intention of the review process to nit-pick a protocol to death nor to induce a hostile atmosphere. However, new investigators need to recognize that the HUC will help the investigator meet all ethical and regulatory guidelines before the protocol can be approved. Protocols which are not clinically and scientifically sound are, by definition, a waste of the subject's time so can not be approved. Investigators should anticipate having to make minor changes to the consent form and, potentially, modify their protocol as a result of the review process.

2. Every protocol will be submitted in on electronic media as specified by the Director of Research. Every protocol must be in the format specified by the institution and all human subject concerns must be addressed within the protocol and consent form. The protocol must be complete to be considered. This means that all consent forms, signatures, impact statements, etc. must be present.

3. Every protocol will be due a minimum of three weeks prior to the announced meeting at which it will be reviewed.

4. The Director of Research assigns a minimum of three in-depth reviewers to each protocol. At least one of the reviewers will be a subject matter expert in the area under consideration. At least one of the reviewers will be a member of the HUC.

5. The protocol will be sent via electronic media to all members of the HUC and to all in-depth reviewers. The principal investigator will be informed of the identities of the in-depth reviewers. The reviewers will communicate with the investigator about their concerns via electronic media open to all members of the HUC. Thusly, all members of the HUC and reviewers can keep abreast of the ongoing comments and responses concerning the protocol. In this way, all of the major questions about the protocol should have been raised and answered well before the committee meeting. A revised protocol is normally not due before the meeting.

6. The committee meetings are designed not to be unfriendly, threatening, hostile environments. Rather, the committee members attempt to ascertain any problems with the protocol not identified and corrected during the electronic media review and consultation process. The principal investigator or designated co-investigator must be present either telephonically or in person at the HUC meeting to answer the committee's questions. No protocol will be considered without such representation. The investigator has about five minutes to present a summary of the project and changes made on the basis of communication with the "in depth" reviewers. Flow diagrams, illustrations of key equipment, etc. are helpful in clarifying the presentation. After the presentation, the "in depth" reviewers present any unanswered concerns either in person or by telephone. The floor is then open to discussion of all scientific and human use aspects of the project. A majority of members must be present to approve a protocol unless it is a minimum risk protocol. Minimum risk protocols can be approved when a minority of the members are present as long as a majority of the members have provided written votes in advance of the meeting.

Minimal risk is defined as risks not considered greater than those encountered in the subject's daily life or during routine physical or psychological examinations. Types of studies which may meet this requirement include (a) anonymous interviews and surveys which do not ask personal questions or questions about use of illegal substances, health care delivery / epidemiology reviews, educational methods studies, and (b) collecting data from existing records.

7. The HUC may approve, disapprove, require modifications of, or postpone final consideration of any protocol. Approval is by vote of a majority of members present.

8. The HUC will review each protocol once per year or more often as required by the HUC members in consideration of the level of risk to subjects or to unanticipated events. Protocols are approved contingent upon the investigator's agreement to inform the HUC of any changes in the protocol, discovery of increased risk to subjects, or problems with / injuries to subjects and to provide detailed progress reports at intervals determined by the HUC at the time of approval.

9. An approved protocol can not begin until the corrected final version is handed in to the Director of Research.

10. The Director of Research, acting as chair of the HUC, may convene an expedited review sub-committee of the HUC consisting of at least a scientist, a clinician and a community member to approve minor changes in protocols.

F. Compensation of subjects for injuries: The FDA informed consent regulation (21 CFR 50.2516) requires that for research involving more than minimal risk, subjects must be told whether any compensation and any medical treatment is available if injury occurs and, if so, what it consists of, or where further information may be obtained. The institution's own policies govern eligibility for compensation or treatment. These are not mandated by the law.

G. The HUC has the right to inspect the documents (patient records, data, and consent forms) pertaining to all approved protocols at any time with no notice to the investigators. If the institution is performing any research under the approval of the federal government, this right is extended to representatives of the FDA regardless of whether the individual protocol is related to

FDA regulation.

H. Ethical guidelines: All protocols performed under the auspices of the institution must strictly adhere to the ethical guidelines laid down by federal regulation and international treaties. These guidelines were detailed in the previous chapter (Research Ethics). However, just to reiterate the key point:

Under no circumstances shall the welfare or treatment of a subject be put second to participation in a research study. No patient's treatment will be substantially delayed or denigrated to a clinically important extent due to participation in a study. It is the absolute responsibility of the investigator to protect the protocol's subjects from risk.

I. Consent to participate in a study: Every participant in a research project, to include controls, must give informed consent to participation IAW the institution's regulations. Some minimal risk studies do not require consent forms. These are called exempt studies. You still need to put in a protocol for scientific review but you can request the committee to permit you to give the subjects a written or verbal explanation of the study in lieu of a signed consent form.

J. Maintenance of data and consent documents: Written copies of informed consents must be maintained by the investigator until the protocol has been completed at which time they must be turned over to the HUC for indefinite storage. The investigator must maintain copies of the protocol's raw data for a minimum of 25 years. Get help from your research administrator on this.

Chapter 8

Pitfalls in the first steps

I. Can you trust the interpretation of study results by people who have not adequately tested their assumptions or are not neutral?

Sometimes data are interpreted incorrectly in all innocence or because unwarranted assumptions are made or some vital co-variate is unknown to the investigators. These are pitfalls in interpretation which happen all the time and will always be with us. However, many organizations refuse to fund studies whose results may not accord with their interests. Should such a study be funded and the results do not accord with their interests, the results may be suppressed by using restricted publication rights incorporated into many grants. Should the results be published, they may be distorted in the public press by ignorant / biased / self-serving politicians and writers. See section J for examples of both problems.

II. The following are errors in experimental design and performance everyone tends to make that you do not have to repeat:

(Modified and extended from (loosely based on) the "Handbook in Research & Evaluation by S. Isaac and W. Michael, 1971, Robert R. Knapp of San Diego)

A. Common errors in formulating a research study

1. Put off selection of a problem until there is not enough time to plan and perform a reasonable study.
2. Uncritically accept the first research idea you think of or is suggested to you.
3. Select a problem that is too vast or too vague to investigate meaningfully.
4. Prepare fuzzy or untestable hypotheses.
5. Fail to consider methods or analysis procedures in developing a tentative research plan.
6. Fail to get adequate advice from experts in every field related to the study (e.g. a physician failing to work with a psychologist when using a stress questionnaire as part of a study).

B. Common errors in reviewing the literature.

1. Carry out a hurried review in order to get started. This usually results in overlooking previous studies containing ideas and methods that could improve the study.

2. Believe that a computer search can actually identify most of the important articles. Do not use known references to find others.

3. Concentrate on results section rather than evaluating and using valuable information on methods, measures, etc.

4. Do not look beyond journal articles for information; e.g. do not talk to published experts.

5. Rely on local clinical opinion instead of getting help from experts in the area.

6. Fail to define the breadth of information so conduct too narrow a search which results in missing vital background clinical and methodological information.

C. Common errors in research design and methodology

1. Fail to define the research population.

2. Use a sample too small to permit analysis of sub-groups or to tell the difference between the main groups. ***** This is the most common problem in clinical studies !!!! Not enough subjects participate to be able to tell the difference between groups even if there was one!!!!*

3. Attempt to perform a study in one year that requires many.

4. Fail to plan data collection procedures so that they are accurate and workable.

5. Fail to perform a pilot study to test the design and instruments.

6. Experimental and control groups are not selected in an unbiased way.

7. Insufficient subjects to prove the hypothesis.

8. Use subject as own control when treatment causes a permanent change.

9. Attempt to match subjects on irrelevant or too many variables.

Practical exercises for section A

1. Can the double-blind, placebo controlled protocol presented as sample two at the end of the clinical research book be performed ethically? Give a yes or no answer and support it with your logic and information provided in the ethics section of this book or other sources. Be sure to define the concepts wash-out periods, placebo, single blind, and double blind and explain how they relate to the design of this study..

2. Read the protocol using non-human animals presented as sample four at the end of this book.

a. Are you convinced that this needed to use non-human, rather than human, subjects? Explain why you feel that the crucial information could not have been gotten from post-surgical human patients.

b. Why do you think the study contained a section evaluating the efficacy of ultrasound in determining how completely the wounds healed?

c. Are you convinced that only the minimum number of animals would be used?
Why?

d. Should the animals have been euthanized at the end of this protocol? Why or why not?

e. Does the background section provide the information you need to determine whether this study is needed at all or whether the investigators known enough about the state-of-the-art in ultrasound imaging of incisions to actually perform the study? Detail your conclusion.

3. Are the outcome measures in the protocol on mulehauler's syndrome presented in sample three appropriate for measuring what the investigators want to know? State an explicit set of outcome measures which will answer the question.

4. Use the Internet (Medline and Psychinfo – not just a search of the WWW) to do the following background searches: First decide on a very limited topic you want to search (such as diagnosis of reflex sympathetic dystrophy with thermography). Before starting the search you must have in your possession at least two articles very specifically on that topic. Do the search and see if both turn up. List the articles you started with then state what data bases you searched and attach a list of the articles which turned up. How did you do?

5. At what point in the development of an innovative technique do you believe:

(a) it needs to undergo formal evaluation for efficacy.

(b) it can be incorporated into your regular practice.

6. Psychotherapy has been proven ineffective as a cure for cramping phantom limb pain. Assuming that psychotherapy is within your scope of practice:

- (a) Can you use it to try to cure cramping phantom limb pain?
- (b) Can you be successfully sued for doing so?

7. If you are an experienced educator with no clinical credentials or clinical training:

- (a) Can you legitimately do EEG biofeedback with normal school kids whom you are assigned to work with in order to give them a “mind-extending” exploration?
- (b) Can you do EEG biofeedback with children diagnosed as having ADD to decrease their symptoms?

8. Make up a clinical idea you might wish to test.

- (a) State its hypothesis in a non-specific / non-testable way and in a specific/testable way.
- (b) Use the examples of limitations on feasibility provided in this section to make up your own list of limitations for your study.

Section B

Basic study structures for the office and clinic environment

Chapter 9

The logic and progression of designs

A. Basic ideas:

1. There is a logical progression from "pilot" style clinical case studies through lots of steps to comparative, controlled studies. I'm going to describe typical study designs in several ways so you are going to see the same information in different contexts. This isn't to bore you but, hopefully, to give you an idea of how the various designs are used and where they fit into the flow of clinical research.

2. Studies can be classified by what they are attempting to find out:

- > Exploratory studies - investigate novel ideas.
- > Confirmatory studies - replicate or confirm exploratory studies.
- > Explanatory studies - work out the details of the above.

3. The type of study you choose is determined by:

- > what you want to know and
- > what your resources are.

B. The special case of pilot studies: When you have an idea for a new treatment, technique, etc. you usually have little idea of whether it will really work. You may have tried it with a few

patients as a best clinical bet but you probably have no way to estimate variability of response. Obviously, you can't estimate how many patients you will need to differentiate between groups or pre-post measures unless you can estimate variability. You also need an opportunity to test your data gathering techniques to determine whether they are reliable across raters, measure what you are interested in, and are actually doable in your environment. All of these factors are worked out during a pilot study. This is a mini trial which lets you test your methodology to see if it can work and gives an opportunity to train your colleagues and work out the bugs. There will be bugs.

C. Definitions:

(adapted from Land and Chase, *Optometry & Vision Science* 67& 68, 1990 & 1991)

Most studies are either experimental or observational.

1. Experiment: A planned study in which the effects of different levels of independent variables are observed on a quantity of interest (the dependent variable).

2. Controlled experiment: An experiment whose design features two or more treatment groups (one of which is a control group) into which the subjects are assigned randomly.

3. Observational study: An investigation in which two or more treatment groups are to be compared but in which the assignment of treatments to individuals is not made by the investigators. For example, two different treatments might be in common use at one hospital and which a patient gets is determined by the preference of each patient's health care provider. This is a type of study where the investigators do not manipulate the variables - they just watch and evaluate. This does not mean that excellent pre-treatment and post-treatment baselines can not be taken.

4. Retrospective study: An observational study in which subpopulations that are identified on the basis of the outcome factor or dependent variable (e.g., has or does not have the condition, died or recovered, etc.). These groups are sampled and then compared to determine if there are differences between them .

5. Prospective observational study: An observational study in which subpopulations that are identified on the basis of an input factor under study are sampled to form the treatment groups. These groups are then followed, evaluated, and compared with respect to the outcome factor under study.

There is a "shorthand" method for identifying study designs which has become virtually universal so you need to be able to recognize it. Here is an example:

A – B – A

A = initial baseline

B = intervention

A = post-intervention baseline

So, what does this mean?:

A – B – A – C - A

D. There is a logic to the order in which studies are performed

1. Designs: The study design chosen depends on what you want to know and where you are in the process of establishing a techniques validity.

Single subject designs: The earliest design is the single subject case study. Here is where the clinical concept gets its first test. This design is also used to test outcome measures, data gathering techniques, and interventional methodologies. Good records are kept and the best baseline and follow-up possible in the clinical environment are performed. Each subject is considered “unique” and patients are treated differently as the technique is developed. Thus, no attempt is made to group the data by means, etc.

Single group studies: This design is similar to that of the single subject study but all subjects are treated as similarly as possible and everything that can be standardized is. Excellent clinical data gathering, baselines and follow-ups of appropriate length, and standardized treatment protocols are emphasized. The data from subjects can be combined for statistical evaluation to get an idea of response variability, etc.

Controlled studies: In this design, similar subjects are appropriately divided into several groups. If there is a standard treatment, one of the groups generally receives only the standard treatment while another group receives both the standard and the novel (experimental) treatment. It is unethical to withhold a treatment known to be effective if such exists.

Placebo controlled studies: When doing a study, it is frequently worth evaluating the effect of providing an intervention which the patient believes to be efficacious but really is not. This permits estimation of the amount of change due to “non-specific” effects such as patient-therapist interactions, belief in the intervention, the therapeutic milieu, etc.

Placebo effects will be discussed in greater depth later.

They can be very powerful and may account for 50% or even more of a drug, surgery, or behavioral intervention’s effect. It is now considered appropriate to maximize this effect to optimize the entire treatment’s overall effectiveness.

Correlational studies: These studies relate change in one variable to change in another. For example, among young school children foot size tends to be highly correlated with spelling ability. However, an increase in foot size probably does not cause the increase in spelling ability. Nor does increased spelling ability cause increased foot size.

A frequent mistake in clinical studies is to mistake correlation for cause. Several examples of this error will be presented as we go through the evidence supporting the relationship between

change in physiological systems and change in symptoms for a variety of disorders.

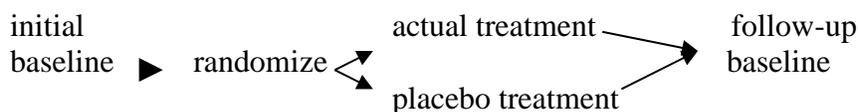
2. Progression of typical designs:

a. **Single subject** - A-B-A design to do an initial test of a new intervention: A very good diagnosis is established and the baseline and follow-up are long enough to account for variability in the disorder and collect data often enough and in enough depth so you know what is happening to the patient. But each subject is different (you don't know just what type of people with the disorder the treatment will work on) and you keep tweaking the new intervention to optimize it. You can not combine data from subjects to make averages so many case studies are reported.

b. **Single group**: A-B-A design used to more carefully test a new intervention after the single subject design shows the treatment at least produces some change for some people. It has the same design as the single subject design but tight entrance criteria to reduce variability in subjects (e.g. for wound healing only take one age range as wound healing rates change with age) and keep the technique the same as possible. You can combine data from subjects to make averages.

c. **Two group controlled study**: A-B-A design again but this is the first attempt to either (1) compare the efficacy of the new treatment with a standard treatment or (2) sort out the non-specific effects of the new therapy from the actual effect of the therapy. Non-specific effects include the placebo effect, patient-therapist interaction effects, etc. The experimenter and the subject usually don't know which group the subject is in (single and double blind designs).

The typical design looks like this:



Randomization has to be real and not contaminated by the experimenter or a technician to avoid bias.

Blinding the study:

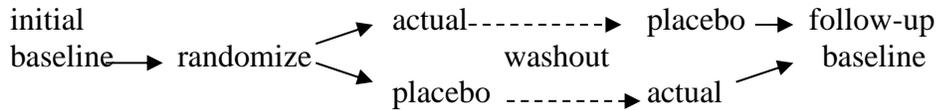
Single blind - The subject does not know which group he/she is in but the people doing the evaluations do.

Double blind - Neither the people doing the evaluations nor the subject know which group the subject is in. Crucial to avoid bias.

Placebos: The placebo needs to be believable and realistic or it is useless.

(e.g. biofeedback placebo of EEG which can't be felt or of an irrelevant muscle which doesn't effect the problem but which patients think should effect it). See Pollard and Katkin (1984) for an example of biofeedback placebos. Remember that you need to be certain that the "placebo" is not an effective intervention (e.g. problem with Kegels).

4. Cross over design: Same type of design as the two group controlled study but the subjects eventually receive both treatments as follows: A-B-A-C-A
 Baseline - randomization to get treatment B or C, get the treatment, permit a “washout period” so the effect stops, change to the other treatment, final follow-up.



When can't you use a balanced cross-over design??? You can't use this design if the treatment doesn't stop - e.g. teaching somebody to relax. If it works, they won't stop using it!
 The space between the first and second “treatment” contains a wash-out period so the effects of the first treatment end.

1. If the effect of one of the treatments doesn't wash-out within a reasonable period of time.
 (e.g. study on muscle tension recognition training for low back pain among soldiers going into combat)
2. If the effect of the real intervention is so clear during the treatment phase that subjects are convinced that they had the real treatment and won't waste their time crossing over.
 (e.g. failed initial double-blind PEMF headache study)

E. Phases: (I admit this is a bit repetitive but it comes from a slightly different direction.)

1. Logical flow:

a. Exploratory pilot / clinical feasibility - first human application of the idea in which the viability of the concept and its safety are assessed to at least some extent. In this stage, you give your best clinical try at helping the patient with your idea. The effective part of the therapy gets partialled out in later studies which are complex enough to look for potentiation of parts of the therapy, etc.

b. Clinical research - the idea is evolved and perfected in well defined groups of patients. includes large single group studies, controlled studies, and comparative efficacy studies.

c. Clinical validation - multicenter studies assessing the technique's effectiveness and safety on large numbers of subjects.

d. Post acceptance studies - long term follow-up studies of large numbers of patients undergoing the procedure.

2. FDA style phases for investigating new procedures (nearly always drugs):

a. Phase I - Try the drug on healthy people to see if it does any harm and to establish a safe dose.

b. Phase II - Try the drug on a group of patients with the disease it is aimed at to see if they do better than *retrospective* controls.

c. Phase III - Large scale randomized, controlled trials to determine if the drug works better than placebos given to matched controls and how it compares to other matched patients given other drugs used for the same purpose.

Obviously, the "FDA" style series of phases is not appropriate for many surgical and behavioral interventions.

3. Multicenter outcome trials when (a) not enough subjects are at one center and/or (b) or "real world" evaluations need to be done.

F. The design depends on what you want to find out:

1. Evaluate the effectiveness of a therapeutic approach - This requires a progression from exploratory single group studies through comparative studies which nearly always have matched placebo controls.

2. Demonstrating the validity and reliability of a new diagnostic test - This requires a cross-sectional approach in which the results of the new test and a gold standard are compared.

3. Demonstrating the validity and reliability of a screening test for picking up a disease in a general population - Similar to testing a diagnostic test but requires a far larger sample.

4. Identification and reliability of prognostic indicators - Longitudinal cohort studies are used to figure out which risk factors are associated with the onset of which problems.

5. Cause and effect - Finding out if a change in one variable causes changes in an other variable is the most difficult because simple correlation is not enough. A combination of longitudinal studies with multiple observations and controlled studies must be used.

G. Longitudinal vs. Cross-sectional studies (regardless of whether they are observational or experimental):

1. Longitudinal studies - determine changes in a group or groups over time. they can be either prospective or retrospective in nature.

a. Prospective (cohort) studies begin with variables and determine the outcome. The main types are:

(1) Deliberate intervention (have one group of low back patients use exercises while the other group does not).

(2) Observational (just sit back and watch what happens as with a group at high risk to develop low back pain - see if they do).

b. Retrospective (case-control) - start with an outcome and determine the variables that determined it. The main types are:

(1) Deliberate intervention (was there a difference in return of back problems between a group treated with exercise vs. one treated with surgery).

(2) observational (why did some subjects develop back pain while other, apparently similar, subjects did not?).

2. Cross-sectional studies - describe one or more groups as they are at one point in time.

3. Outcome: A clinical outcome study is exactly what it says - a study specifically designed to determine how well a clinical technique worked. There is nothing special or spectacular about them. It's more an idea/concept of what the clinical investigator wants to accomplish than a special design in itself.

Outcome studies range from single group designs (intended simply to determine whether a technique worked at all) to multicenter, multi-treatment, double blind, controlled studies (intended to determine which of several treatments and placebos are most effective and cost least). these cost-effectiveness studies are normally far beyond the realistic abilities of small clinical groups to conduct.

H. Choosing the design:

1. The design you choose is partially determined by what are you attempting to accomplish and the "stage" you are at in demonstrating what-ever point you are attempting to make. There is no one way to look at experimental designs. There are too many types. many are meant for very special purposes. The major types commonly used in the clinical setting are presented below. The study you want to do may not fit into any nice category. The idea is to be as logical as possible.

2. When using experimental designs (which are all prospective), there is a logical sequence to establishing the effectiveness of a new technique. This sequence applies to most clinical research questions.

a. Case studies are where the practitioner tries a new idea based on previous clinical and/or research findings. Good history, diagnosis, outcome, and follow-ups are done. this permits refinement of techniques and determination of whether the technique is actually any good upon follow-up. The case study is very close to good clinical practice.

b. A group of similar patients with similar conditions treated with fairly

standardized technique to see if the technique is generally effective. the single group design can include taking extended baseline and follow-up measures. It can also include internal control periods where each patient gets the treatment for part of the time and not for other parts of the time. The results of your study would be compared with the results for the general population as reported in the literature.

c. Controlled outcome study done with several groups. placebo and control groups are likely to be introduced at this stage to differentiate (a) placebo and (b) "course of time effects" from effects of the technique. When possible, neither the investigators nor the subjects know which group a particular subject is in or which treatment is being received (double blind). These studies are likely to have cross-over types of designs so that all subjects get active as well as placebo treatments during the study. This strategy precludes subjects who need a treatment from not getting it.

d. The last step is frequently the comparative study in which the technique is compared with others also used to treat the disorder. This study is frequently part of a major clinical trial of the technique done at many centers. in this type of design, several similar groups may be compared.

I. Reiteration of the need for realistic controls in studies of technique effectiveness:

1. 35% cure rate for chronic pain patients with placebo medication.
2. The effect could disappear shortly after end of treatment.
3. Apparent changes could be caused by habituation, biorhythms, halo effect, etc.
4. Effective portions of complex treatments determined.

Vertosick (1998), gives a chilling rendition of what happens when the wrong controls are used. He reviews the twenty year life of a surgical technique called the extra-cranial-intracranial bypass. This technique was used with patients until a few years ago who were at high risk for having strokes. Blood from the scalp was shunted into the brain through the skull. The developer of the technique did a small study in which he compared the survival of his patients to historical controls gleaned from the literature. His patients survived better than those found in the literature. The technique caught on among neurosurgeons and dozens of specialized laboratories opened to teach the technique. The insurance companies jumped on the bandwagon and the technique was big business. Nobody demanded controlled studies before using the technique on thousands of patients. Eventually a placebo controlled study was performed. It turned out that the technique was even worse than helpful. It did harm. Lots of harm. Within a year or two, the insurance companies stopped paying and the technique, along with the expensive training laboratories, virtually died out. What went wrong? First, the historical controls turned out to be quite dissimilar from the originator's patients! The next thing, and worst, event was that

physicians accepted a poor study as proof that the technique was effective and never checked their own results!

J. Assignment to groups: Both conscious and unconscious bias are very likely in subject selection. For example, an independent investigator charged with assigning subjects to groups may not want to place an especially sick subject into a placebo group. Thus, assignment must be by some blind process. Subjects can be assigned at random, by matching, or by stratified then random assignment (e.g. all divided into degree of severity of disorder then randomly assigned to groups within categories).

K. Doing a study without impacting the normal treatment process: We have two main ways to try new treatments without interfering with treatments already used for the problem:

1. Perform the new treatment while patients interested in participating in the study are on a waiting list to receive the standard one: A good example occurred during development of a new treatment for Morton's neuroma of the foot. This is an annoying, painful thickening of tissue between bones controlling two toes. It usually develops slowly and, when recognized early, is not debilitating. Most patients with the problem can still walk and do their jobs so the surgery is not an emergency for most of them. When this study was done, the waiting time for this type of surgery for non-debilitated patients was about two months. Patients diagnosed as needing surgery for the problem were put on the waiting list but were also told that they could have the experimental treatment while waiting. If it worked, they could skip the surgery. If it didn't they would still get the surgery on time. The experimental treatment was a series of several injections intended to reduce pain from the area. It turned out to work better than the surgery for most patients. The few who did not improve with the injections had their surgeries on time.

2. Divide patients interested in participating in the study into three matched groups and give (a) one group the standard treatment, (b) the second group the standard treatment as well as the experimental one at the same time, and (c) the third group the standard treatment and something that looks and feels just like the experimental treatment but doesn't work (a placebo). This can only be done when the standard and experimental treatments do not interfere with each other. The effectiveness of the new treatment is determined by whether more members of the group receiving it get better faster than do members of the other groups.

This technique was used to try to find ways to help some people with severe asthma to use less steroids without reducing their comfort. Interested asthmatics were divided into three similar groups. One group was given their usual treatment, one was given their usual treatment in addition to an anti-inflammatory drug called Methotrexate, and a third group was given their usual treatment plus a placebo that was similar to Methotrexate. Most members of the group getting Methotrexate could take less steroids without loss of comfort than could members of the other groups.

Chapter 10

Exploratory single subject and single group designs

A. Single subject designs - the exploratory study:

The idea of the single subject design is to look at the impact of one or more interventions on changes in a disorder's activity on each individual participating in the study. The data from many subjects are not combined. Rather, individual differences in reaction to the intervention(s) are evaluated on a case by case basis.

The importance of single subject designs to clinical progress can not be understated. When a clinician has an idea or hears about a technique which seems worth trying, careful planning needs to be done so the clinician can really tell whether the technique had any effect. It is largely through exploratory trial of new ideas that we make initial progress and set the direction for where productive clinical studies can head. Of course, the technique to be tried has to meet the ethical standards for trying it on a patient - especially including having a reasonable chance of producing as much or more success balanced with correspondingly less side effects and intensity of intervention as would be available using accepted techniques.

When faced with having to make efficacy decisions based on only one or a few subjects, it is crucial to know as much about that (those) subject(s) as is possible. You don't have the luxury of hoping that differences will balance out between groups or that the majority of subjects will show improvement. Instead, you need to maximize the information you can get from your subject. That means truly understanding the natural history and expected variability of the disorder and combining that knowledge with your patient's personal history.

1. Baselines: The key to successful single subject designs is the longitudinal baseline which extends sufficiently before and after the intervention so you can tell whether the disorder's activity changed. You need to follow the subject long enough and take measurements of the disorder frequently enough so you are confident that you know the variability of your patient's problem. The "guesstimated" length of the pre-intervention baseline is set by expected variation in the disorder. For instance, if you are working with classical migraine headaches and the literature says that they can be expected to occur an average of one to four times per month, then you can not establish a baseline level of activity in less than three months. However, your patient may report having consistently had two or three headaches per week virtually every week for the last four years. In that case, a one month pre-treatment baseline would do. The real problem is that many chronic problems vary tremendously over the course of a year or years due to hosts of concurrently changing variables including seasons and other medical problems.

Occasionally this is simply impossible. For example, cluster headaches which occur in

apparently random discrete bunches once or twice per year can not be fit into this pattern. You simply have to give your trial but have to constrain yourself from deciding whether it worked until your follow-up has been long enough to cover the period during which several clusters would have occurred. This circumstance is incredibly difficult because you don't know whether to try the method on more subjects or put it on a mental shelf and just wait. I try to take care of this problem by being very selective about who I try the idea out on. For relatively rare events such as cluster headaches, I find patients who have long histories of having attacks at least every three months. These patients are more difficult to locate and the technique precludes just trying the idea on the first patient that walks through the door, but produces useable results in a useful period of time.

The length of the post-intervention (follow-up) section of the baseline is set not only by the issues of variability in disease activity which controlled the length of the pre-intervention baseline, but by how much you want to know about the duration of your intervention's effectiveness. The later is very different than the former because knowing that you had a temporary/initial effect on the disorder's level of activity tells you little about the actual value of the intervention as a therapeutic approach. The usual approach is to see if there is any positive effect. If there is, the subject is followed closely to get an idea of any long term effects which may occur but the initial change is used to justify trying the intervention with several more subjects. The crucial point is not to confuse short term changes with demonstrations of clinical efficacy.

When you carefully control both the length of the baseline and the frequency of readings (data points), you are attempting to pin down the pattern and characteristics of, rather than simply control for, what is called the unsystematic variance or error variance in the observed changes. This is the variance due to uncontrolled fluctuations of your outcome measures as opposed to systematic variance which is the change due to your intervention. In a typical controlled study you would attempt to control for error variance with matched control groups so the changes will, hopefully, happen about the same in all groups and potentially cancel each other out. The combination of adequate time and multiple recordings during the single-subject study's longitudinal pre-treatment baseline gives you the potential to determine what is causing some of the variation for an individual subject so its effects can either be better predicted or controlled. For example, it may become apparent that every time the humidity reaches a critical threshold, your subject gets a worse migraine than usual. You would carefully track humidity and reduce the reported intensity of the headache proportionately when analyzing the data. It might also become apparent that eating cheese causes some of your patient's migraines. You would have the patient either eliminate cheese or have it at a consistent rate and amount (dosage) to reduce its apparently random effect on headache occurrence. Of course, you need to minimize unsystematic variance as much as possible in any design by carefully selecting subjects who have the minimal number of impinging factors changing randomly and by trying to insure that the subjects' environments and clinical situations remain as similar as possible throughout the course of the study.

The strength of single subject designs is really that they recognize that every patient is very different and will respond very differently to an intervention. They permit the investigator to look for the differences and work with them rather than trying to submerge them in a huge pool of data. This strength is especially important when a new idea is being tried out because the

intervention is likely not to be in its optimally effective state and may need considerable tweaking as it progresses. Thus, the single subject design is far more likely to pick up a weak effect than a controlled study because it is not contending with having to eliminate the pooled variance of the group from the apparent effect size. The weakness is that very few subjects can participate relative to the amount of time and effort that goes into conducting the study so the results are less generalizable to a varied population.

2. Changes in variables to look for: There are three major changes in disease activity which need to be assessed during the pre and post-intervention baselines.

a. Intensity of the disorder. For instance, are the headaches less severe after intervention?

b. Rate of change in the disorder: Acute disorders with recognized progressions in intensity may show a change in rate of progress. For example, the rate of change in intensity of cold symptoms can be predicted so an alteration in that rate is evidence of a successful intervention.

c. Variability in intensity and frequency of occurrence of episodes or in changes in intensity.

3. Common designs: Numerous designs are used to strengthen single subject data.

a. ABA (baseline, intervention, baseline): This is the most common and practical design when conducting an exploratory clinical study. It is frequently used with surgical and behavioral interventions because the intervention tends to be very long lasting or permanent when it works so there is no way to get it to stop on demand.

b. ABWBA (baseline, intervention, washout period/baseline2, intervention, baseline): This design is used to strengthen the conclusion that the variability in the outcome measure is due to the intervention and to help determine how much of the variability actually was due to the intervention. However, it can only be used with treatments which stop working in a reasonably short period of time. This is fine for short acting drugs with brief washout durations.

c. ABWCA (baseline, intervention 1, washout period/baseline2, intervention 2, baseline 3): This design is used to compare two drugs (or two dosages of the same drug) which have a reasonably short washout period. It is also used to give a placebo behavioral treatment before a (hopefully) real behavioral or surgical intervention. It can be used to give a treatment which does have a washout period prior to one that does not - as when a drug is compared to a surgical or behavioral intervention. This design permits using the power of a placebo or comparison without resorting to a controlled study. You can make this incredibly complex by giving the first intervention again after an additional washout period to strengthen your evidence that the intervention controlled the change in the outcome measure.

d. AB¹AB²AB³A (baseline, intervention trial 1, baseline, intervention trial 2, baseline, intervention trial 3, baseline): It is occasionally possible to give discrete trials of a treatment, such as training sessions, in which it would be expected that the training would have an effect on the outcome measure for a while and then drop off but that there would eventually be a cumulative effect. This is used to demonstrate a stronger relationship between the intervention and the observed change than could be done if the intervention was just done once or if the entire training period was considered the intervention.

B. Single group designs - the case series / clinical replication:

Clinical replications are an excellent way to begin showing that an intervention is effective for groups of people. Because data from many subjects can be combined, an idea of the overall effectiveness of the technique can be ascertained.

The designs discussed above for the single subject studies are also used for single group studies. The single group study is very similar to the single subject study with three crucial differences:

1. The subjects are kept as similar to each other in type of disease and confounding factors as is possible. Thus, if age is important to a wound healing study, you would only use a very small age range. You would set inclusion criteria which minimized participation by anybody that had concurrent problems which would add to variability in the outcome variable(s). Thus, no diabetics would be able to participate. This means that the intervention's "generalizability" to the overall population of people with the disorder can not be tested directly.

2. The intervention is standardized and written down carefully so it can be applied with the minimum amount of variation required by differences between individual patients. The histories of both surgical and behavioral interventions are filled with instances of supposed "case series" and "clinical replications" in which the power of using a group of subjects is entirely lost because the intervention techniques were radically different for each subject. This happens especially frequently when clinical investigators at several sites "pool" their data.

3. The data from all subjects are pooled to give an estimate of variation in effectiveness. This permits shortening the baseline periods. The availability of pooled estimates of variation permit the use of inferential statistics to (1) derive confidence limits, (2) determine whether it is likely that the pre to post-intervention differences in the outcome measures are due to chance alone, and (3) to perform a power analysis on the data to determine how many additional subjects would be required to test the hypotheses further using the same or other designs to some degree of probable significant difference.

Chapter 11

Observational Studies - Longitudinal & Cross-sectional

A. Strengths and weaknesses of typical observational study designs:

(Much of this information was compiled by Hulley and Cummings, 1988.)

Observational studies are those which do not manipulate any variables. They are designed to follow the "natural history" of a disorder. For example, following all members of a graduating class for ten years to determine who gets sick is a longitudinal cohort, observational study. It is observational because the investigators do not intend to do anything to the class (or part of it) during the course of the study. It is longitudinal as the subjects are followed over time and it is a cohort study because all members of the group of interest are followed from its inception. This is crucial as it avoids missing data from those who drop dead (etc.) before you get the study going.

1. Observational longitudinal studies use many measurements over time to track changes in a population. It is crucially important that they start with every member of the group of interest (the inception cohort) because of the bias that creeps in when members of the original group who become sick or die before the first measurement disappear and are, thus, not accounted for. As the disappearance may relate to the intervention being studied, the success rate may be far lower than shown by the study. The reverse is also true as healthy people are less likely to answer surveys and are far more likely to disappear from a hospital's ongoing records and clinics so successes may not be counted because they are not contacted.

Longitudinal studies have the ability to pick-up and sort out variation in the disorder due to both extraneous and related factors. They require less subjects than cross-sectional studies because more of the variation can be accounted for by repeated measurements. However, these are usually long, expensive studies which require great effort to keep in frequent contact with the subjects or so many disappear that the study's validity disappears.

a. Prospective Cohort: This version of the longitudinal observational design is frequently used to determine either (1) incidence of a problem in a population or (2) failure rates of interventions such as surgical procedures when little is known about the subjects before they arrived for the procedure or when the procedure was for an acute problem such as a traumatic event so a history of variations in the outcome variable are not relevant. For example, if you are studying the effectiveness of a new method for acutely pinning traumatically broken collar bones, there is no pre-intervention history of changes in collar bone healing to be recorded for the vast majority of your population.

The prospective nature of the design has the tremendous advantages of giving the

investigators control of the type and quality of the measurements and of insuring that the investigators can determine whether the subjects actually meet the inclusion criteria. This is the optimal design for establishing incidence rates because all subjects were followed from the start.

b. Retrospective cohort designs (the chart review study): This design is used when the intervention or event of interest took place some time ago and patient records must be used to find out what happened. These are shorter, less expensive studies than the prospective designs and frequently comprise the bulk of required "resident" research projects. They have the distinct disadvantage of having to count on medical records, etc. which may not be accurate and may have missing information. Crucial patients may have disappeared so these are rarely true cohort studies. There is no ability to control which measurements are made or to check the appropriateness of the subjects. No or limited ability to get information about the subjects' pre-participation backgrounds is possible.

2. Cross-sectional designs:

a. Population (usually people with a particular characteristic or disease): One measurement is usually made of many people over a brief period of time. Mail response surveys intended to establish prevalence of a disorder or correlations between the disorder and other variables of interest are examples of this design. Large numbers of subjects are required because there is no way to account for much of the variability in the measure of interest due to extraneous variables. This design usually can not establish cause and effect because it has to count on unrepeated correlational data. This design is usually of briefer duration and lower expense than longitudinal studies. The problem of non-respondents is substituted for the longitudinal study's problem with dropouts.

b. Retrospective case-control studies: These usually consist of one measurement of a group of people who have the problem and a (usually) matched group of people who do not have it. You start with people who are already sick or not so do not have to wait to see who gets sick. Differences in risk factors are easier to pick up than in the other designs. Random sampling is very difficult so bias is a weakness. Huge sample sizes may be needed to pick up rare events.

B. Regression to the mean is a real problem in evaluating the effectiveness of a purported treatment because patients tend to enter when they are sickest so tend to get better over time.

C. Establishing cause and effect:

1. Longitudinal study with many data points to account for natural variability.
2. Change the purported cause in effected subjects and observe the effect.
3. Establish a dose-response relationship.
4. Remove the cause from unaffected subjects and see if incidence decreases.
5. Several controlled studies need to be performed - not just one.

D. Epidemiological studies (from junkscience.com 2003):

1. Case-Control Studies: An observational study where the researcher starts with subjects

with the disease of interest and examines their history to see whether an exposure of interest is statistically associated with the disease of interest.

2. Cohort Studies: An observational study where the researcher starts with disease-free subjects and follows them into the future to see if those with the exposure of interest get the disease of interest.

3. Ecologic Studies: An observational study where data is collected on populations rather than individual subjects. (*The*) Researcher associates differences in disease rates between geographically distinct communities with some "exposure factor".

Chapter 12

Prospective experimental studies

A. The concept: As opposed to observational studies, experimental designs involve purposely manipulating a variable and then determining the result of that manipulation. Normally, levels of the important variables are recorded both before and after the manipulation. Thus, virtually all experimental designs are prospective in nature.

B. Typical Designs:

1. **Single subject:** One or more subjects are followed very carefully over time. Variability in outcome measures is evaluated on a subject by subject basis. These "within subjects comparisons" are very powerful ways to find out if a new idea has any merit and to track changes in a very rare disease where only a few subjects are available. This design was discussed in detail in the preceding chapter on exploratory studies.

2. **Single group:** Similar to the single subject design but pre-to-post changes can be compared using between subjects analyses. This design was also discussed in detail in the preceding chapter on exploratory studies.

3. **Prospective longitudinal:** This design is very similar to the longitudinal cohort studies discussed in the preceding chapter on observational studies except the observations begin sufficiently long before the intervention so the outcome measure's variability can be ascertained with considerable certainty. Pre-intervention to post-intervention comparisons can use multiple or single subject statistics. Repeated measures and time series analyses are usually used to differentiate random changes from changes related to the intervention. These designs are frequently used when small groups of patients are available and when the duration of effect needs to be determined. The details of establishing adequate pre-treatment and follow-up baselines have been discussed elsewhere.

4. **Parallel group:** Two or more groups each receive a different treatment and differences in the result are compared. One of the groups may receive a placebo. This would make the study a "placebo controlled" design because there is a control group which should show changes only due to factors extraneous to the active component in the intervention.

Subjects may receive a third, standard treatment in addition to the novel intervention and the placebo so the actual groups would be (a) standard plus placebo and (b) standard plus novel. This avoids having to deny subjects a treatment known to be at least somewhat effective while a new treatment is being tried. Of course, this approach can not be used if the therapies conflict. The appropriate performance of these designs while subjects are on a waiting list for a surgical procedure meant to correct the problem or while on a baseline for the standard treatment are discussed elsewhere.

Whenever possible, the team evaluating the outcome should be entirely blind to which treatment the subjects received in order to avoid bias. The patients should also be blinded as to which treatment they received to minimize expectation effects. The study is "single blind" if either the subjects or the evaluators know which treatment the patient received. It is "double blind" if neither know. When all these factors are in place, this design becomes the famous "double blind, randomized controlled trial".

Parallel group, paired comparison group, and longitudinal studies are all used to test the relative efficacy/effectiveness of several treatments over time. Changes over time are frequently important to ascertain because one intervention may be more efficacious initially but be of far shorter duration and survival times can be significantly different. In other words, a slight increase in efficacy may not balance a significantly decreased life expectancy.

5. Paired (matched) comparison: Pairs or triplets of subjects are matched as closely as possible on all variables likely to effect the outcome and are then usually randomly or sequentially assigned to one of the intervention groups. Each group receives a different treatment (or an active treatment and a placebo treatment) and the results from each group are compared. The groups are usually generated using stratified random or stratified sequential sorting from a single subject pool. They usually use the same control and evaluation strategies as discussed for the parallel design discussed above.

6. Crossover: These designs include several different treatments or a treatment and a placebo. Subjects are randomized so that each could get either the placebo or the real treatment first. The first treatment is given for a set period long enough to get an effect and then stopped. The second treatment is then given after a sufficient period of time has elapsed so that the effects of the first treatment dissipate. Obviously, this design can not be used if the effects of the treatment do not stop after a reasonable length of time. Differences in response due to treatment presentation order effects are evaluated before the main analysis is performed.

7. Run-in: This is a single-blind predecessor section added to any of the controlled study designs and can be thought of as being the appetizer before a main dish. It is used to reduce variation in response to the main intervention or reduce the number of subjects required. For example, a placebo can be given for a few weeks to find out who will be compliant with the study instructions or drop out. A low dose of the active intervention may be given to eliminate those who overreact or do not react at all. Obviously, the investigators know that the patients are in a run-in portion of the study but the patients think they are in the first part of a multi-treatment design such as a crossover study.

8. Comparative / cost comparison: These complex studies use parallel/independent style designs to compare a novel treatment with an accepted one to find out which is more effective and less expensive.

9. Factorial: This complex design is set up to measure the effects of several variables simultaneously and is unlikely to be used by office based clinicians. A simple 2 X 2 factorial design would be set up as outlined in Table 5.

Table 5 A typical 2 X 2 factorial design study

	placebo	active 1
active 2	placebo + active 2	active 1 + active 2
placebo	placebo	placebo + active 1

C. Why go to all the bother to conduct a randomized, double blind, placebo controlled trial?: A properly designed study of this type provides an excellent evaluation of the effect of an intervention (relative to that of a realistic placebo) on a well defined group of patients.

1. The prospective nature of the design permits the investigators to entirely control how the data are gathered, which patients are included and which instruments are used. Quality control can be very good. This helps insure that all appropriate patients are included in the randomization process.

2. Reduces the likelihood of bias by using blinded, neutral evaluators.

3. Permits differentiation of effects of the active intervention from those of the placebo - which, in turn, eliminates expectation, therapist - patient relationship, and other effects from the analysis.

D. So, when don't you do a double-blind, randomized, placebo controlled study? Placebo controlled studies are crucial to most, but not all, demonstrations of technique effectiveness. For example, the study on a new fixation technique for clavicles discussed at the very start of the book, described using a device which had already been shown to be safe and effective on many bones with a different bone for essentially the same purpose - holding the ends steady and compressed while healing takes place.

The "new" element was that nobody had performed this kind of fixation with this bone to prevent malalignments which resulted in lumps under the skin. The first study was a pilot to make sure the technique worked and that the collar bones remained stable and lump free for a few months - long enough for the known length of normal healing to be over with. The second phase was to try the technique with a hundred or so patients who were followed longitudinally both clinically and radiographically to establish failure and complication rates. The third stage will be to monitor thousands of patients receiving the technique from many surgeons to look for long term complication and failure rates. So, why no placebo control? We already know that when healthy people's collar bones are left to heal on their own, they do so after a while but in whatever position they happened to settle down in. There are no other techniques which attempt to adjust and then hold the clavicle in a normal position so there is nothing with which to compare it. There is no purpose to performing a placebo surgery as the amount of lump and its interference with function is objectively measured.

However, let's say that the aim of the procedure was to change subjective comfort rather than function. Then you might compare people having a standard procedure (usually a sling) with those having the surgical procedure. The sling would be a poor control because the surgery requires a much larger investment on the part of the patient. First, a surgery has to be endured and, second, the patient needs to walk around with a metal device sticking out of their collar bone for a few weeks instead of just an innocuous sling. Thus, the proper control would be to do the procedure but leave the collar bone misaligned. Then the investigator would have the patients rate their comfort levels every month or so for a year to give time for the placebo effect of the extra effort to wear off. I wouldn't do this study. Would you?

Here are other specific instances where a double-blind design would be skipped:

1. When not providing the innovative treatment might lead to irreversible harm. For example, single group studies may have provided convincing evidence that a novel intervention for a rapidly fatal disease is highly efficacious in extending subjects' life spans with no decrement to life style. The investigators would conduct large open trials of the intervention and compare the effects with historical data for standard treatments or with survival data from other geographic areas not yet using the innovative treatment.

2. When subjects can not be randomized because of ethical or other problems.

3. When insufficient subjects are available to produce a meaningful result. Far too many studies are published in which insufficient subjects participated for any difference between the groups to be detected. It may be that intersubject variability is so high that more subjects would be needed than can be gathered. In this case, a prospective longitudinal design is probably more appropriate. This concept was detailed in the section on initial power analysis.

4. When sufficient funds and/or time are not available to perform the study properly. If you can't wait for an appropriate clinical end point or to gather-up sufficient subjects, don't start. If the least expensive method for recording meaningful data is just out of your reach, don't substitute one that really can't give the answer you need and waste everyone's time and money.

E. A word about choosing your placebo: If you are going to use a realistic placebo, be certain it actually isn't an effective treatment on its own. In my first foray into treating exercise induced urinary incontinence among female soldiers (Sherman et al 1998), we used Kegel exercises as a highly believable but "ineffective" placebo control intervention for vaginal sEMG biofeedback. Numerous prominent clinicians led us to believe that Kegals would not be effective. They were. In fact, the placebo control group and the biofeedback groups both did superbly.

Nocicebo (negative placebo) effects are very powerful and have to be noted carefully. We conducted a placebo controlled trial which involved placing a large magnetic field generator adjacent to patients' thighs. A patient who turned out to be receiving the placebo (an inactive generator) happened to have a cramp in her leg while receiving her first exposure. She firmly believed that the device had caused the cramp although she was assured that this was not one of the device's side effects. Her placebo response was very great.

F. How Blind are double blind, “placebo” controlled studies?

The cornerstone, perhaps even the "holy grail" used by medical research in the United States to prove that a treatment is more effective than an ineffective control is the "double blind, placebo controlled" design. For this design to work, none of the participants or anyone associated with the study can know which patients are receiving which treatment. This means that the placebo has to be realistic enough that the patients can not tell that they are getting an ineffective treatment. This becomes especially difficult when a cross-over design is used in which patients sequentially experience both the real and the placebo condition with a brief "wash-out" period in between. If the placebo is a sugar pill with no side effects and the actual pill leaves you sick, dizzy, and flat on your back, every-one will know who is on the real medicine because the sickened patients, who of course have a pretty good idea that they are now on the real drug, complain about the side effects to someone - even if that someone is not an investigator, and the word spreads. This happened in a study using pulsing electromagnetic fields to treat headaches. One therapist collected the data, talked with the subjects, and put them on the devices. The therapist initially did not know which machine was which. After a few weeks of listening to which patients improved and which didn't the secret was out. The problem was solved by restarting the study with two therapists. One put the subject on the machine but was forbidden to find out how the subjects were doing. The other therapist did not know which machine the subjects were on but tracked their headache activity.

The following is a paraphrase of a typical comment about the validity of this design which appeared on the internet psychophysiology chat group in February, 1998: "I have been involved in hospital wards doing medicinal research for over twenty years and after dozens of studies, have yet to see one where all of the nurses and patients didn't know which group they were in."

It is now accepted practice to have subjects rate the likelihood that each condition was the "real thing" just after completing their course. Several studies using the outcomes of these ratings have appeared recently and found that the patients and nurses only thought they knew which groups they were in when, in fact, their guesses were random. However, these studies were examining recent medicinal experiments. See McQuay et al (1995) et al for a discussion of this problem and evidence that high quality “double blind studies actually are blind. Such experiments now frequently use very realistic placebos which have mild side effects of their own.

It certainly makes sense that the design is useless if a believable placebo can not be invented which corresponds to the actual therapy. The placebo intervention has to parallel the actual intervention in every way possible or it won't produce the same effects. Thus, it has to have just as much patient-therapist contact, just as much homework, and just as much “machine time” with just as much reinforcement for whatever the subject is supposed to be doing. Having a control group in which people are on a waiting list or examining ceiling tiles for “awareness” just doesn't work. The sad thing is that investigators frequently rush into using this design prematurely so, in fact, waiting list controls frequently do as well as the actual treatment group. This means that the investigators wasted a fortune in funds, their time, and their subjects' time because they performed this design before performing good single group studies with long enough baselines to

tell what the variations in the disorder's course were.

G. The crucial importance of adequate follow-up periods and follow-up studies: Most longitudinal studies are designed to keep track of their subjects until all of the changes that are going to happen are likely to occur. This is certainly not the case with experimental designs. Many initial studies stop days after the intervention and declare success even though virtually no follow-up information is available.

This practice, and the gradually disappearing practice of accepting such papers for publication, has led to terrible mistakes which do real harm to patients. Examples abound in the history of every area of clinical practice where the failure to wait long enough to find out if the apparent success is real has led to the dissemination of techniques ranging from invasive ones such as open heart massage and microvascular scalp to brain artery transplants to psychotherapy for phantom limb pain. None of these techniques would have gained popularity and held on for so long if adequate follow-ups had been done. The phantom pain literature provides a typical example of this nonsense. At least sixty-three absolutely useless techniques were in use as recently as a decade ago because none (NONE!!!) of the clinical case series upon which the proliferation of the techniques was based followed their subjects for more than a few weeks or, at best for a month or so. None followed their subjects long enough to find out that the pain either never really went away or came back shortly after the end of treatment.

No prospective study should be designed without incorporating an adequate follow-up because there is no way to know whether any observed changes have any meaning.

High quality journals, such as the Journal of Bone and Joint Surgery, now refuse to publish studies which do not have adequate follow-up periods. This practice has considerable promise for reducing the proliferation of ineffective interventions.

Chapter 13

Outcome and quality of life studies

A. Concept: When you hear the expression "outcome study", people are frequently actually referring to studies which go beyond determining whether the intervention effected the specific problem it was aimed at to determine whether doing the intervention actually had any effect on the person. The idea is incorporated in the old joke about the operation being a success but the patient died.

If you are proposing to perform an expensive, complex surgical procedure on an older patient, you want to know if the procedure will actually help the patient. If the patient has a slow growing bone cancer and is very likely to die of some other disease in progress before the effected bone hurts or breaks and interferes with quality of life, there may be little reason to perform the surgery. The same idea would go for a hip replacement if the patient's pain is not significantly reduced and mobility is not significantly increased for a reasonable number of years.

There is certainly a trade-off between a treatment that produces a slight increase in longevity at the cost of quality of life. This trade-off isn't limited to major surgical and chemical interventions late in life. The decision as to whether to deny oneself gustatory pleasures for an entire life by eating in what is the current fashion's "healthy" way has to be balanced with the potential effects on longevity and quality of life in later years. If you are only going to add a few months to your life and quality is not likely to begin to slip significantly until the last year or so, it may not be worth it to give up something you really like.

B. What should be assessed? When you talk about quality of life, what do you mean? What do you want to include? Measures of physical, psychological, and social well-being are certainly core to the issue. Physical well being would include the ability to perform normal activities of daily living as well as desired activities such as walking without pain and loss of breath. The following list is partially based on ideas described by Spilker (1986):

1. Functions of daily life include the ability to bathe, dress, and feed oneself; bowel and bladder control; ability to walk sufficiently to perform necessary chores; and ability to move and bend well enough to get out of cars, furniture, etc.

2. Functions in the work place include the ability to work productively at the chosen vocation and ability to support oneself at a satisfactory standard of living relative to the pre-disease state.

3. Ability to perform social roles and maintain family and community relationships.

4. Ability to perform hobbies and pastimes.

5. Intellectual capabilities including memory, ability to communicate, ability to make decisions, and ability to think, act, and react.

6. Emotional stability and health including mood stability (swings), beliefs about the future, and emotional levels.

7. Satisfaction with life including level of well-being, perception of general health, and outlook for the future.

8. Signs and symptoms of illness including, but not restricted to, the disease being treated. This includes an assessment of the nature, severity, duration, frequency, and impact of the problem and the required treatments.

C. How is quality of life assessed?

1. Numerous “standardized” tests are available which purport to assess quality of life. Most of them are so general that they are nearly useless for evaluating patients with a particular problem. However, they do provide a set of questions which have been well validated on various populations. Adapting these tests by using appropriate questions (with permission) is a good way to increase the validity and reliability of the instrument you eventually develop. The □paper and pencil□ inventories seem to emphasize subjective measures (how the patient feels about life and what the patient feels can be done) while the □in-clinic□ evaluations usually emphasize some combination of physical abilities such as how long the subject can walk on a treadmill at some speed without pain or running out of breath (or some objective change in respiration). The □bibliography on Health Indexes□ lists the current crop of inventories (US Public Health Service).

2. When you are attempting to select a measure of quality of life, check that:

a. The measure was tested with and makes sense for your type of patient and disorder. It must test the parameters impaired by the clinical problem you are assessing.

b. The measure can actually pick up changes in the parameters you are interested in with enough sensitivity and reliability that you can see changes from your intervention. Too many of the tests are so global with such broad categories of intensity responses that only huge changes in life could be picked up.

c. The measure's test-retest reliability is high enough for the magnitude of change you expect to be detected and not overwhelmed by random fluctuations.

3. Examples of quality of life scales (mostly based on information from Pynsent et al, 1993):

a. Quality of Life Index (as summarized by Pynsent et al 1993 from Spitzer 1981): This scale was developed to measure overall well-being of cancer patients. Quality of life is assessed by scoring five qualities on a scale of zero (not functioning) to two (OK) . The characteristics of life required to get each number for each quality is explicitly and clearly

defined so any two people scoring a patient are very likely to come up with very nearly the same score (interrater reliability ranges between 0.74 and 0.88). The qualities are intellectual activity, daily living, subjective assessment of overall health feelings, support system, and outlook on life. The maximum score is ten. Unfortunately, the scale is so general and has so few numbers that people have to be very sick to score far below ten and huge changes have to take place before they would be reflected in a significant increase in numerical rating.

b. Nottingham Health Profile: (developed by Hunt et al 1985 and described by Pynsent et al 1993): This ten minute questionnaire contains 38 questions about energy level, pain, emotional reactions, sleep, social isolation, and physical activities. It has been used extensively with British orthopedic patients (hip replacements, fractures, arthritis) and has good reports of test-retest reliability (0.88) and validity relative to physical exams.

c. Arthritis Impact Measurement Scale (developed by Meenan et al 1980, described by Pynsent et al 1993): This self-administered scale takes about fifteen minutes to complete and is very appropriate for patients with musculo-skeletal disorders. It consists of 45 questions which ask about mobility, physical activity, dexterity, household activity, social activity, activities of daily living, pain, depression, and anxiety. Its test-retest reliability is 0.90 and has been well validated with many hundreds of patients having arthritis but (so far as I know) not other conditions.

d. Sickness Impact Profile: (developed by Bergner et al 1981 and described by Pynsent et al 1993): This inventory takes about a half hour to either administer as an interview or complete as a questionnaire and produces both physical and psycho-social sub-scores. It contains questions on sleep, eating, work, home management, recreation and pastimes, ambulation, mobility, body care, movement, social interactions, alertness, emotions, and communication. It has a test-retest reliability of about 0.87 when done as a questionnaire and 0.97 (!!!!) when done as an interview. It has been validated for use with several types of orthopedic patients.

e. Barthel Index (developed by Mahoney and Barthel 1965 and described by Pynsent et al 1993): This scale was developed to assess chronic musculo-skeletal and neuro-muscular problems. It rates the categories of feeding, transfer from wheelchair to bed and back, personal toilet, getting on and off the toilet, bathing self, walking on a level surface, going up and down stairs, dressing, and bowel and bladder control. Its test-retest reliability is 0.89 so it is reasonable to use for patients who have severe disabilities. The Index of Independence in Activities of Daily Living (Katz and Akpom 1976) is quite similar in outcome.

f. PULSES profile (Moskowitz and McCann 1975): This is among the oldest and best known scales so many people have been trained in its use. Four levels of impairment are defined for physical condition, upper limb functions, lower limb functions, sensory functions, excretory function, and mental/emotional status. It has an incredible inter-rater reliability of 0.95 and test-retest reliability of 0.87. Once again, it takes a great deal of change in the patient to see much change in the scale.

g. Quality Adjusted Life Years: (developed by Rosser 1990 and described by Pynsent et al 1993): This scale combines the value of extra years of life with increased quality of life provided

by an intervention. Several groups have added information about monetary cost of the intervention vs. cost of the changes in life style. The scale rates both subjective distress ratings and objective disability ratings. This is a complex scale which measures very complex concepts. I urge readers to get reviews of the scale before attempting to use it to rate the monetary value of life.

D. Specialized outcome scales: Many specialties have developed detailed scales for the specific disorders they deal with. These scales are frequently very well validated and should be at least looked at before an investigative team tackles developing a scale from scratch. If nothing else, the team can get an idea of what has been tried and where the weaknesses are. For example, Pynsent et al (1993) have produced a book which contains chapters on standardized ways to assess many of the areas evaluated by orthopedic researchers including the spine, shoulder, elbow, hand, hip, knee, foot, and ankle. Each chapter contains details on the various scales used to assess the part in question along with a discussion of why the part should be evaluated in particular ways. The American Academy of Orthopaedic Surgeons (1994) has also published a book on outcome studies entitled "Fundamentals of Outcome Research" which contains considerable information including ways to use data from outcome studies to decide whether to change clinical practice.

Chapter 14

The protocol's research plan and design

A. The protocol as the first part of an article: When you write a research protocol, you are really writing the abstract/summary, introduction/background, and methods sections of the article which should eventually grow from the completed study. Thus, you need to write clearly and in enough detail for somebody who is not an expert in your field to follow exactly what you are saying and understand what you want to do and why. The reference citations are interspersed in the introduction and methods section exactly as they would be in a typical clinical research article so readers know how you are substantiating the assertions which underlie your premises.

B. Functions of the literature review: Your literature review should lead the reviewers directly to the question you are going to pose and then answer in your protocol. They should finish the introduction being able to anticipate the next logical step in clarifying whatever problem you are working on - and they should find that question to be the one you are asking. A literature review is not a list of articles into which you interject a few comments such as "___ et al (1602) found that _____ happens." You need to give enough of the results of the most important studies so that reviewers can tell how the prior study supports your proposed work. This is especially crucial when the study you are noting appears to have tackled the same question you are asking. You must make it clear what the limitations of the previous study were and how your study will counter those weaknesses.

The other information supplied by the introduction is the rationale for your methodology. Reviewers need to be convinced that the methods are reasonable and measure what you are interested in. This is where you relate your outcome measures to the problem you are investigating.

C. The method section: This is the most detailed section of the protocol. The method section has to be sufficiently detailed and clear that a reviewer could replicate your work from reading it.

Many method sections are so outlined and cryptic that the rationale for the methods are not clear and the study design gets lost in the details. Reviewers are helped by having:

1. A brief overview of the design/method section.
2. A flow sheet which summarizes patient participation and evaluations.
3. An outline of the experimental design including a box with the groups outlined in it.
4. Diagrams of all unfamiliar equipment and techniques.
5. Copies of all logs and questionnaires which are not well recognized standards.

Differences between the various groups in the study have to be made very clear - usually with the help of the outline noted above. The outcome variables and exactly how you will measure them have to be very clear. This includes giving the reviewers an idea of what types of

numbers or answers you are likely to get.

The statistics sub-section is one that most investigators try to ignore but is the one on which reviewers - especially grant reviewers - spend considerable time and concentration. The reviewers need to be convinced that the investigator has some idea of how each hypothesis is going to be tested. This means that you need to relate each hypothesis to specific outcome variables and then to the way you will determine whether the data you gathered supports or does not support the hypothesis. A well written statistics section begins with a brief overview of the logical approach and then lists each hypothesis or question with the methods for evaluating and testing it detailed. This certainly includes specifying statistical tests and justifying your choices. This is the place you set your acceptable levels of significance and justify them.

Sample Study

Look at the introduction to the pilot study:

Does the literature review support using pulsing electromagnetic fields to prevent onset of migraine headaches? Does it provide a rationale for why the technique might work?

Look at the introduction to the controlled sample study.

Does the literature review give better support and rationale than the pilot study did?

Does the literature review support using a one month baseline? Does it support the need for a control group?

As a knowledgeable clinician, are you convinced that the investigators have supported the chosen outcome measures?

Does the literature review provide enough diagrammatic material for you to be comfortable that you know what a pulsing electromagnetic field device is and what the log the subjects will use is like?

Now look at the method sections for both protocols.

Can you tell what the investigators plan to do? Are there enough flow sheets, diagrams, etc. to make the plan clear?

In the full protocol, are the differences between the groups clear?

Can you tell what the placebo is? Is it believable?

In the full protocol, is the statistics section clear? Does it address each hypothesis or question?

Chapter 15

Defensive reading of clinical literature - does the design fit the needs?

A. Does the design make sense?

1. Can you tell what the study design is? I don't mean a special name (e.g. crossover control), but, rather, what the authors did.
2. Is the design consistent with the stated purpose of the study?
3. Is the design appropriate given the state of knowledge about the topic or the current understanding of research design?
4. Is the design too complex or inadequate for the purpose?
5. Are threats to the validity of the study identified? Are they corrected where possible?
6. Are potentially confounding (extraneous) variables controlled by the basic design?
7. Is use of the outcome measures supported by the literature review?

B. The instruments (devices, surveys, etc):

1. What instruments were used to measure the outcomes / concepts?
2. Were they adequate reflections of the outcomes being studied? If they do a great job measuring something not related to what the investigators wanted to know, they are useless.
3. Did the investigators show that the instruments were appropriate to measure the required information?
4. Do the instruments measure accurately enough for the required precision?
5. Are the validity and reliability of the instruments adequate for their application? What

problems would be expected with the selected instruments validity and reliability?
How were they addressed?

6. Were the instruments appropriate for the population being studied?

7. If the instruments were developed for the study, what were the procedures used to assess their adequacy?

C. Procedure:

1. What specifically was the treatment?

2. Did the investigators present convincing evidence that the treatment should effect the disorder in question?

3. What procedure was used for data production? Is it clearly described? Were the procedures appropriate? Could the methods have influenced the findings? How?

4. Could another investigator repeat the same study, given the description of the procedures?

D. The diagnostic criteria: You must be able to tell what the diagnoses of the participants were. The diagnostic criteria must be sufficiently clear so that you could select the same type of subjects and so you can compare the authors' results with results of other studies.

Chapter 16

Pitfalls in study design

A. Errors in outcome study design and performance everyone tends to make which you do not have to repeat:

(Modified and extended from (loosely based on) the "Handbook in Research & Evaluation by S. Isaac and W. Michael, 1971, Robert R. Knapp of San Diego)

1. Common errors in formulating an outcome study
 - (a) Select a problem that is too vast or too vague to investigate meaningfully.
 - (b) Prepare fuzzy or untestable hypotheses.
 - (c) Fail to consider methods or analysis procedures in developing a tentative research plan.
 - (d) Fail to get adequate advice from experts in every field related to the study (e.g. a physician failing to work with a psychologist when using a stress questionnaire as part of a study).
 - (e) Fail to determine the outcomes of critical importance.

2. Common errors in outcome study design and methodology
 - (a) Fail to define the patient population adequately.
 - (b) Use too few patients to permit analysis of sub-groups or to tell the difference between the main groups. ***** This is the most common problem in clinical studies !!!! not enough subjects participate to be able to tell the difference between groups even if there was one!!!! *****
 - (c) Attempt to perform a study in one year that requires many.
 - (d) Fail to plan data collection procedures so that they are accurate and workable.
 - (e) Fail to perform a pilot study to test the design and instruments.
 - (f) Experimental and control groups are not selected in an unbiased way.

- (g) Use patient as own control when treatment causes a permanent change.
- (h) Attempt to match subjects on irrelevant or too many variables.
- (i) Choose measures not capable of detecting the information required.

3. Common errors in gathering data

- (a) Pay insufficient attention to establishing and maintaining rapport with subjects. This often leads to refusal to cooperate or to a negative attitude that can reduce the study's validity.
- (b) Weaken research design by making changes for administrative convenience (this is a very common problem!!).
- (c) Fail to explain the purpose and tests to colleagues who will impact on subject availability and attitude.
- (d) Fail to evaluate available measures thoroughly before selecting one. This often leads to the use of invalid or inappropriate measures or ones with such low validity that the study is wasted.

4. Common errors in observational studies

- (a) Observers insufficiently trained to give identical ratings.
- (b) Observations required are too detailed or time consuming for the observers to actually gather in the time they have available.
- (c) Fail to insure that the observer's presence will not alter the activities being observed.
- (d) Attempt to evaluate activities that occur so infrequently that reliable data can not be obtained.

5. Common errors in the study of relationships:

- a. Confuse correlation with causation.
 - b. Use correlational tests for independent data on data which include variables normally correlated with each other (e.g. height and weight among children). This invalidates the analysis.
 - c. Use simple correlation techniques in situations requiring partial or more

complex methods.

d. Use parametric correlative techniques for non-parametric data.

e. Use the shot-gun technique of correlating everything with everything else and then report spurious "significant" correlations.

6. Common errors in retrospective research

(a) Insufficient chart and literature data available to conduct a worthwhile study or test the hypotheses.

(b) Poor definition of study population.

(c) Data analysis not integrated with the literature.

7. Common errors in descriptive research

(a) Objectives not clear thus tons of useless data collected but never used.

(b) Sample selected on the basis of convenience, not randomly.

(c) Analysis not planned until data are collected.

(d) Obvious biases in the data collecting devices so that subjects do not give accurate answers.

B. Examples of typical problems with experimental design and evaluation: (Very loosely adapted from *Principles and Practice of Research: Strategies for Surgical Investigators* by H. Troidl et al; Springer-Verlag, New York, 1991.)

1. Lack of comparison with a control group or the literature: *"Closed repair of fully disrupted trigger finger pulleys among twenty traumatically injured, young, otherwise healthy males resulted in successful return of normal motion for all but 4.3 percent who required open revision of the repair. Significant morbidity occurred in only five of the cases. Thus, we recommend adoption of this technique as the standard of care."* □

There is no way to know how these results compare to the expected levels of success and morbidity for this problem. The authors should have compared it with the results of standard procedures. They certainly would have to conduct a reasonably large comparative study giving their treatment a □head-to-head□ comparison with randomly sorted subjects before they could recommend replacing other techniques with theirs.

2. Improperly matched groups: *"Distance to walk on a treadmill without pain was compared for a group of soldiers with tibial stress fractures who were treated with pulsing magnetic fields (PEMF) for two weeks with historical controls treated with non-weight bearing."*

The group treated with PEMF could walk significantly further than the controls so it is concluded that PEMF is an effective intervention for this problem. □

The problem here is that the historical controls included all comers with tibial stress fractures while the experimental group consisted of soldiers. The soldiers were very likely to have been younger and in much better shape than whoever the historical controls were. Motivational factors also have to be factored out in studies such as this because compliance with usual concurrent treatments is greatly effected by how motivated individuals are to get going. Thus, the soldiers were far more likely to improve more quickly.

Practical exercises for section B

1. Describe the following experimental designs (including types and durations of baselines, typical groups, diagnostic assessments, subject selection criteria, etc.), their strengths and weaknesses, and when in the course of demonstrating the validity of a new idea they would be used:

- a. Single subject design.
- b. Single group design
- c. Multiple group controlled design without crossover
- d. Placebo controlled study with crossover (include the requirements for an excellent placebo group which controls for all non-specific effects and explain how your group will do so)
- e. Longitudinal prospective study

2. With regard to sample protocol 2: Given the results of the pilot study, could the investigators used a typical crossover study design in which subjects were randomized to receive actual or placebo exposure first and then crossed over to receive the other exposure condition after a wash-out period? Would this have strengthened the credibility of the study?

3. Get and read the following paper (I may be able to supply copies - check with me.): Spiegel D et al: Effect of psychosocial treatment on survival of patients with breast cancer. Lancet 2(1989) issue 8669, 888 - 891, 1989. Critique the design of this study (especially contrasting the difference between ways of designing studies which relate results in the experimental group to either a control group or the general population) .

4. Critique the experimental design of the article in Sample 5.

5. Get and read Peniston, E & Kulkosky, P: Alpha-theta brainwave training... in alcoholics. Alcoholism: Clinical and Experimental Research 13, 271 - 279, 1989 (I may be able to supply copies - check with me.):

- a. Critique the original article's design.
- b. Perform a computer based literature search (using data bases such as Psychinfo and Medline) to find other articles by the authors on this topic. State which articles you found. Critique the treatment and the support provided for it by the entire set of articles.
- c. Use a computer based search to locate articles by other authors commenting on the original article. State which articles you found. Critique the critiques of both articles.

Section C

Establishing the credibility of data and clinical publications

Chapter 17

Subject selection techniques - sampling, inclusion - exclusion

A. Importance - Sample bias kills studies: How important is it to get your subject selection system correct? Ask the people who did the huge, expensive telephone survey which predicted that Landon would beat Roosevelt in the 1930s. They did a technically good job but forgot that most of the household phones were in relatively wealthy people's homes. At that time there was a higher correlation between wealth and voting Republican.

B. What to watch for:

1. You need to have a superb grasp of what factors influence your patients' progress. Typical factors include (a) severity, duration, and reversibility of the problem being studied, (b) age and sex, (c) variability in response to the technique / device / medication being tested, (d) concurrent treatments in progress, and (e) changes with time. Changes in your subjects' conditions can be altered by all of them.

2. You have to have several objective ways to measure the changes in which you are interested. It is very rare that one method is sufficient as each tends to have its own

limitations.

The outcome variable(s) must measure the problem you are attempting to influence!!!
An unbelievable number of studies are published in which the outcome measure is not related to the problem. This frequently happens because the authors have not taken the time to read the details on the test they are using as an outcome measure.

C. Types of Variables:

1. Independent variables are those that can change on their own. This does not mean that you can control them. For example, time is an independent variable. However, you do control many of them. Examples include dose of medication you choose as part of the study design, duration of exercise, and which group a patient is placed into. **DO NOT** count on many variables which appear to be independent to actually be independent. You will be surprised by what turns out to influence supposedly independent variables.

2. Dependent variables are those whose values are influenced by changes in other variables. For example, number of bacteria on the skin should be influenced by amount of washing within narrow limits. Many study designs are interested in altering the value of an independent variable and determining the effect on a dependent variable. For example, giving a drug and measuring changes in severity of the disease being studied.

Note that there may be no single, simple way to measure the severity of the disease so there are likely to be a complex of interrelated "dependent" variables.

Are the dependent variables you choose to measure and record your outcome measures?? Not necessarily. Many changes in patients' conditions may be affected by your study design and need to be tracked to see whether they are altering your results.

3. Confounding variables are frequently those that change beyond your control (time of day, concurrent treatments, hospital food, patient - health care provider interactions, etc.) but which alter your outcome measures - usually in complex, varying ways. These are the factors which can destroy your study!!! A few will always sneak through and not be measured until it is too late but you must minimize their effects on your data by recognizing and measuring the effects of the most important. This is one of the reasons believable placebo controls are so important. The changes you observe may be due to changes in a variable you don't even know exists.

D. Sampling: This is the concept of how you choose participants, how you sort them into groups, and how many you need to perform the study.

1. Inclusion - exclusion criteria: It is crucial to eliminate as many confounding variables as possible by limiting who can be in your study. For example, if you are

studying the effect of exercise on wound healing rates, you need to eliminate everybody with problems which would interfere with healing (e.g. brittle diabetes) and either limit your population to one age group or design the study so that sufficient people from the major age groups are included so the effect of age can be factored out of the results.

Exclusion criteria eliminate people who meet the inclusion criteria but have some overriding problem which would invalidate their data.

2. How often the event you are measuring occurs - rare events vs. outliers:

If you are doing an initial study on the effectiveness of a drug at a given dose, you need to be aware of the few people who will be hypersensitive and very insensitive to the drug at that dose. The "rare event" of a patient having serious complications from a drug at a given dose (such as death) needs to be looked for in certain types of trials. If your sample size is too small to pick up rare events, you can not report that the drug is safe at that dose.

When you are looking at your results, there will almost always be a few values that are simply orders of magnitude different than the rest. You can occasionally discard them because you know they are out of range of the test used to produce them.

Unfortunately, this is rarely the case and you have to report them as you can not determine whether you are looking at a true mistake or a rare event - outlier which is a portent of troubles ahead for mass use of the intervention.

3. Sampling techniques:

a. Judgmental: You decide which patients should be in the study. As patients who meet your criteria come along, you make up your mind whether you think they should be in your study. Don't do this.

b. Convenience: Grab the folk who happen to be in the waiting room when you decide to try out a survey. Don't do this either unless you are just trying something out in very rough form.

c. Consecutive: Take every patient that meets your criteria as they are identified. This is the safest method for a clinical study because there was no way for anyone to influence who got into the study.

d. Random: Random sampling is the best way to choose subjects for a study from a large, apparently homogeneous population. This works well for yeast in a well mixed vat and inbred mice in a huge colony but rarely works for clinical studies because finding a large homogenous population is nearly impossible when working with real diseases. Factors such as age, severity and longevity of the disease, and gender are frequently overwhelming variables that need to be controlled for so actual random sampling is rarely used. Unfortunately, most parametric inferential statistical tests require random sampling from a homogeneous population as a crucial entrance criterion to use the test. So, many of these tests can not be used to assist in understanding the outcome of clinical studies unless their sample sizes are very large.

e. Systematic: If you have a list of thousands of patients who probably meet your criteria, selecting every "n"th name is very acceptable. Choose "n" so that you wind up with the number of subjects you want. For example, when I was surveying veterans who were amputees, the US Veterans Administration supplied an alphabetical list of 25,000 names and addresses but wanted to survey only 5,000 amputees. I chose every fifth name to increase the odds that there would be no bias in selection.

f. Proportionate: A random sample can be disastrous when used with a non-homogeneous population. If you are doing a study in which age may be important (as when healing rates are a factor) and you are working at an institution that sees a disproportionate number of young adults (as would be the case in a military hospital), a random selection would result in under-representation of older and very young age groups. You would have to select more from these age groups or change your entrance criteria so only young adults participate in the study. This situation is very common when a large group of people are selected by a mail response questionnaire in which race may make a difference. Our society being what it is, races are not randomly distributed throughout a geographic area and location covaries with race and economic status (e.g. a disproportionate number of Blacks live in poor neighborhoods). Sampling a proportionate, representative sample of middle class Blacks means identifying and sending a sufficient survey to this sub-group which means that your survey is disproportionate.

g. Stratified Sequential or Random: You must assign patients to groups without bias. The optimal way to do this while obeying the entrance criteria for most statistical tests is random sampling. Unfortunately, our patients are so different from each other in their histories and likely responses to whatever treatment (or etc.) you are testing, that you probably can not sort them into groups randomly without increasing the variability of your results to an unmanageable size. For example, if (a) you only have a few appropriate patients available to study, (b) age and degree of disease effect your results in a very important way, and (c) most of the young, relatively healthy patients happen to wind up in one group, the results will not be valid. If the groups are different initially, the difference can easily be artificially magnified by the study.

(1) Factors to sort on: The best factor to use is the one whose change is most likely to skew your data. If two factors will have important effects, then sort on both. It is very difficult to sort on more than two factors so find the one or two that make the most difference.

(2) Procedure: Let's say that you want to study the effects of three drug dosages on some disease but you know that both sex and youth vs. old age have an overall effect on the drug's functioning. You also know that you only have a few subjects available so you can not just run lots of people and try to factor the effects of age and sex out at the end. You could do two variable, stratified, sequential random sampling into three test groups (e.g. stratify by two age classes <young & old> and sex <m & f> into three dosage groups (low, med, & high)). Follow the flow chart below when going through this example. Assume the first subject who has met all of the entrance criteria is a young female. First put the person into the correct age category (young). Within the age category, put the person into the correct sex category (female). Next, either sequentially or randomly assort the person to one of the three dosage groups using any of the standard techniques such as last digit of the social security number matched with a

random number table. The example in Table 6 shows ten sample subjects and Table 7 shows them sequentially sorted into dosage groups.

Table 6

Ten subjects to be sorted for a study

Subject #	age	gender
1	young	female
2	young	male
3	old	male
4	old	female
5	young	male
6	young	male
7	young	male
8	old	male
9	old	male
10	young	female

Table 7

Stratified/sequential sort for the above ten subjects

Sorting Factor												
Age	Young						Old					
Gender	Male			Female			Male			Female		
Dosage Group	L o w	med	high	l o w	med	high	low	med	high	low	med	high
Sort (subject number)	2	5	6	1	10		3	8	9	4		
	7											

E. Checking the sort: Once you have selected your subjects, check to be sure they really are reasonably random and, if groups have been selected, that the groups really are similar on the crucial variables. For example, in the study of veteran amputees, we checked State resided in for both the people selected to receive surveys and of the entire population of amputees to insure that the proportions remained constant because we felt that climate might be a factor in intensity of stump pain. You can check sorts between groups for randomness with an appropriate statistical test ("t" or "u" for two groups, ANOVA for 3 or more).

F. The art, science, and politics of recruiting subjects:

If you can't get enough of the right kind of subjects, you can't do your study. You also can't coerce people into participating. Even if you don't have any ethical scruples about coercing subjects into participation, you need to recognize that these participants will not provide accurate data and are likely to be as uncompliant as they can get away with. So, what do you do to get enough genuine volunteers if they simply don't fall into your lap? The most common answers are (1) widening the pool of subjects and (2) advertising. You can widen the pool of subjects by relaxing your entrance criteria (which usually leads to more problems than it was worth) or by adding other pools to draw from by including other practitioners and institutions in the study. Adding other pools to yours tends to work very well as long as you can keep firm control of the intervention and evaluation processes. As soon as widening the pool means adding more practitioners to crucial evaluation processes, you are in a bind because of validity and reliability problems which need to be addressed - usually in a very time consuming and not too successful way. So, think twice before you go down the road for more patients. You may wind up with a poor study because you couldn't control your new co-investigators appropriately.

Advertising brings in a set of people who differ substantially from the general population of people with the problem you are working with. You also need to deal with the masses of desperate people who call in with problems peripheral to the one you are recruiting for. Numerous studies have shown that "normals" recruited from advertisements - aren't. So, if you are going to advertise, have the resources ready to screen respondents very carefully and expect to spend a disproportionate amount of time turning away a disproportionate number of potential subjects.

When participation is demanding, inconvenient, or anxiety provoking for subjects, recruiting is more difficult. For example, if subjects have to (1) make twenty extra trips to the hospital spread over the course of six months, (2) get their blood drawn at each visit, and (3) have to keep a complex daily log for the entire period of participation, recruiting people will be difficult and getting them not to drop out will be nearly impossible without some form of compensation. The best compensation is treatment for the problem they came in for in the first place. Treatment studies have far less recruitment problems than non-treatment studies. Even treatment studies have significant problems getting patients to continue complying with post treatment logs, etc. through the follow-up period. The best solution is to pay people to complete participation in the study. The amount has to be sufficient to compensate for expenses and inconvenience but not enough to induce people to participate for the money. This is unwarranted coercion. The problem is that some people may enter the study in order to get what appears to the investigators to be a trivial amount of money and then don't give honest answers because they initially exaggerated their problems. You also need to have a source of funds from which to pay the people. Grants being what they are, this can be a very serious problem.

Chapter 18

Hardening subjective data

(capturing the elusive wild datum in its natural habitat / unsubstantiated phenomena made substantial)

A. Solid data - the basis of research: "Data" rarely - if ever - exist as undeniable, solid bits of reality. Almost every bit of information to be collected has some slop in it; some room for inaccuracy and doubt about its validity and its meaning. Many times there is considerable doubt about the relationship of the □datum□ to whatever we are trying to find out. Sometimes there is no way at all to objectively measure what you are investigating - such as pain intensity, existence of flying saucers, and distribution of sea serpents. Figure 2 illustrates the fluidity of typical clinical data which have to be solidified enough to use in a typical clinical study. The most common reason studies fail is that data are not gathered accurately and consistently. One of the most common reason studies fail is that data are not gathered accurately and consistently.

Figure 2 The elusive wild datum in its natural habitat



B. Why are we so concerned with solidifying the data? The bottom line is that outcome measures which are not crisp and clear lead to false conclusions.

1. Objectivity is simply not a human trait. There are too many well known instances of unintentional bias altering the results of studies. Thus, you need to spend much of your planning time (a) developing and (b) testing ways to objectively record relevant changes in your subjects.

When methods of taking data are designed, the major effort must go toward distinguishing interpretive from objective data. For example, when observing an interaction between two patients, you could report either "the patients were fighting" or the patients repeatedly came close to each other and backed off. The second way permits you to record actions rather than interpretation of actions. This will permit further evaluation of the patients' activities.

"You can fool all of the people some of the time, some of the people all of the time, but you can't fool all of the people all of the time."

Note that the group you CAN fool all of the time is mainly composed of those trying to do objective clinical research. We are just not watching for our subjects to purposely fool us. But they do it all of the time for a wide variety of reasons ranging from trying to tell us what they think we want to hear to being passive aggressive. Spend much of your time finding ways to ensure that your observations are accurate: check for subconscious and conscious faking at least three independent ways.

The common occurrence of unconscious faking and unconsciously biasing the data which results in self-delusion of a discovery is very difficult for clinical scientists to accept. Everyone recognizes that the innovator has considerable emotional attachment to the innovation and wants it to succeed. But, it is very difficult to believe that the attachment is so great that the average innovator would unconsciously skew the data to support the intervention's efficacy. This common occurrence is termed "pathological science". Rousseau (1992) points out that the widespread image of the scientist painstakingly obtaining objective data, testing every side of a question, and disregarding personal interests is not correct. He notes that Locke recognized the problem in the 17th century and that people have not changed since.

Numerous major fiascos have occurred in science because investigators failed to maintain objectivity and check their results. Well known examples from the basic sciences include polywater - a supposedly polymerized form of water which turned out to be regular old water contaminated by salt and organic compounds and cold fusion - in which a fusion reaction was thought to take place at about room temperature. In both cases, the proponents failed to make the most basic checks of their equipment. In the first place, very obvious contamination was ignored by numerous theorists busy writing about the implications of polywater in spite of publications showing that the findings were all due to contamination. The proponents continued to claim that only the detractors' water was contaminated and that their's was pure in spite of the same spectra being given off by both. The ideas finally died after the weight of negative evidence was simply insupportable and nobody would give further heed to increasingly outspoken proponents.

As difficult as it is to bring an end to delusions in the basic sciences, it is next to impossible to do the same in our clinical arena. At least the basic sciences can show such objective findings

as spectra and readings on voltmeters. They can also frequently control many of the variables which are related to the effect they are studying. We clinicians have to make do with changes in patients' reports of pain intensity and the like to judge the success of interventions working upon ill defined problems. In short, our waters can be far murkier than those of the basic scientist.

One of the major problems in the clinical arena is that therapists tend to remember their successes and forget their failures. Sometimes they are not even aware that they have failed with virtually all of their patients having some particular disorder because their intervention resulted in a very temporary decrease in symptoms.

Beyerstein (1997) has reviewed the types of errors and biases which can lead practitioners and their patients to believe that ineffective therapeutic approaches are actually working. His point centers on the idea that objective procedures have evolved for differentiating fortuitous and apparent improvements from causal ones. Failure to follow these procedures causes the misunderstandings. He feels that the problem is multiplied by mistaking temporary changes in symptoms for long term changes in the underlying disease. He points out some obvious problems such as the disease running its natural course about when the treatment begins and spontaneous remissions caused by psychoneuroimmunological effects. Of much more importance is the fact that many diseases are cyclical or have random increases and decreases in symptoms. Examples include arthritis, multiple sclerosis, allergies, and gastrointestinal problems. Temporary decreases in symptoms can be ascribed to a treatment if a reasonably long baseline and follow-up are not performed. A recent discussion on the internet about behavioral treatment of tics related to Tourette's syndrome substantiates his point. Many people were going on at length about success in treating these tics until the head of the Tourette's society came on (indirectly at first) and emphasized that none of the patients had been followed long enough to go beyond the natural period of diminished symptoms which follow an outbreak and that it was during such outbreaks that the interventions were provided. Thus, it was likely that the symptoms had gone away as usual and would return in their due course. He emphasized that he had an entire drawer full of failed therapies which neglected this salient point and had been subsequently shown to be useless.

An other example of the crucial need to understand the disorder is coupled with the need to have a large enough group to truly assess the outcomes. Sampson (1997) reviews Spiegel et al's (1989) early work on groups therapy / support groups for increasing survival among patients with cancer. The control group was too small to have much chance of detecting outliers and was sufficiently smaller than the experimental group (24 to 34) that outliers had a significantly better chance of being picked up in the experimental group than the control group. The investigators did not compare their survival rates to the known curves from very large epidemiological studies but, rather, only to the control group. Sampson states that the finding that the members of the experimental group living twice as long as the controls was an artifact because the members of the control group died earlier than would have been predicted while members of the experimental group survived as long as predicted. The difference was due to a few members of the experimental group living longer than average and a few members of the control group living shorter than average but all within the normal bell curve for survival for their diseases given their starting symptom levels. Thus, the study proved nothing but created a great deal of furor which has yet to die down.

Very similar problems have been demonstrated in inappropriately controlled studies of

acupuncture for drug and alcohol treatment and for pain. Apparent success increases as the tightness of the controls decreases (Sampson, 1997).

Beyerstein also emphasizes that the power of the placebo effect can not be understated. He is certainly correct in this assertion. For example, we can “cure” up to one quarter of patients with chronic migraine headaches using highly realistic placebos (see the discussion in sample protocol two near the end of the book). Of course the headaches come back when the placebo effect wears off. However, excellent controlled studies have shown that the placebo effect can last for up to six months so only a clinician familiar with the literature on placebos would realize that an appropriate length follow-up is eight months to a year. Negative placebo effects (called “nocicebos”) abound in the literature of placebo controlled studies. You can count on anywhere from one to ten percent of patients receiving an inactive intervention to have stomach aches, headaches, dizzy spells, pains, etc. For example, in one of our studies with magnetic fields, a patient receiving the placebo reported a powerful cramping pain in her leg under the site at which the field was being applied. She was seated with her legs up in a comfortable recliner and had no history of leg cramps. The device was immediately checked by medical maintenance personnel who assured us that the device was not putting out any fields what-so-ever and that only a meter and fan were connected to any power. The device itself can not be felt while functioning and does not cause cramps among its users. However, the patient responded to this nocicebo event with the powerful belief that she was receiving the real treatment and her symptoms were reported to decrease significantly for months. This does not mean that the placebo effect should not be utilized therapeutically when appropriate, it simply means that practitioners should know when they are using it. It is known that a realistic placebo causes the release of endorphins so it is not surprising that they are effective for pain control.

Another problem which comes up frequently is the therapist who does his or her own evaluation of success and/ or follow-up by personally contacting the patient. A plethora of studies have shown that the vast majority of patients tell the provider exactly what they think the provider wants to hear for many reasons including avoiding damaging their relationship with the provider. Thus, they have no idea that the patient still has the problem and is going elsewhere for treatment. Practitioners frequently feel that they have such a good relationship with a patient that the patient would never lie to them - but it happens all the time to every practitioner I know of. This effect has been well documented in large systems (such as the US Army and various Scandinavian nations) where patients are restricted to using practitioners within the system and all of their provider contacts are recorded in the same medical record - which can be examined a year or so after the intervention being tested has been completed.

Another point made by Beyerstein is that many complaints are psychosomatic to begin with so any believable therapy which substantiates the problem and gives the patient permission to do without it should have a reasonable success rate. Somatisizers have very real problems - even if they are not initially physically based - and are a very real problem for the practitioner. Some of the saddest cases I have seen are people who were somatisizers who wound up with true, irreversible, disuse atrophy of the limbs.

2. You get what you expect: Over twenty years ago, I had just been transferred to a military base whose main function was testing new equipment. It contained seemingly endless stretches of open land used for artillery shell impacts and testing odd devices. I had somehow

strayed far off the beaten path while driving my wife and young children across the base when the kids excitedly pointed out a jet apparently standing exactly still a hundred feet or so up in the air a half mile or so away. We had been driving along an arrow straight road for miles and the plane was apparently just ahead of us to the left of the road. Of course, being a typically pedantic intellectual, I couldn't resist taking the opportunity to teach the kids about the optical illusion which makes planes appear to be still because they happen to be flying at the correct angle to your line of sight. Nothing I said could convince the kids that it wasn't a small jet standing still just ahead of the car. I, of course, knew that jets couldn't stand still and would never be just a few feet up in the air. It had to be a very large jet far away. Then the road made a ninety degree left turn under the experimental fighter hovering exactly in one place just over the road. Such machines were top secret then so I saw what I expected. Now that my kids are long grown up, I think they finally believe in that illusion, but they have a wicked smile on their faces when we are together and happen to notice a plane going by.

So, finding what you expect is the fate of researchers. You must defend against it endlessly.

3. **You get what your subjects/patients want to tell you:** The vast majority of adults feel a crucially important need to be viewed in a certain way by anyone they come into contact with - regardless of whether they will ever meet that person again. This issue of face, of not permitting oneself to be diminished in someone's estimation, can not be underestimated when eliciting information from patients which the patient might feel could conceivably cause the clinician to view them in a deleterious way. Many men will destroy their knees rather than cut back on their activity level in front of their peers because they would have to admit they have knee pain. This rationale causes the US Army to loose a vast number of macho male recruits every year because they don't come in for treatment early enough to correct simple problems which grow into incurable nightmares by the time they collapse at the side of the road. The same goes for the disproportionally low number of men who come in for headache treatments. Our anonymous, neutral surveys of soldiers show that nearly as many men have incapacitating headaches as females but the males almost never request treatment because headaches are not a masculine problems. Instead, they probably get blind drunk or drugged to get through it. When they are asked if they have headaches during routine physical exams, the answer is nearly always "no".

The concept of people not wanting to put themselves in a "poor light" is illustrated by who reports unusual phenomena. Most of the people who report unidentified flying objects (UFOs) are people who are more highly educated or come from higher socioeconomic backgrounds than their current circumstances would lead one to expect. This well known phenomenon was tested by a group who sent a fake (but realistic) UFO over a small Northeastern (US) city one clear summer evening and then sat back to see who reported it. They knew that many people had seen it because they were watching. But who talked about it? Who actually called the authorities? Virtually only those who felt that they could afford to lose face in front of their neighbors and authorities called in. Thus, do not assume that your patients are going to tell you what their problems are if the patient thinks that admitting to those problems will diminish them in your eyes (Condon 1968).

You can not record self-report data from your own patients - they will tell you what they think you want to hear. You only think your patients are being open with you. Too many studies

have proven this to be a false perception on the part of health care providers. It is nearly impossible for any practitioner to believe that he/she is not trusted by his/her patients enough for them to be open about crucial information. The thought seems to go: "*Sure they would lie to Dr. _____.* *So would anybody. That turkey is as cold as a clam and doesn't listen to anybody - but not me, I'm so open, warm, and friendly everybody tells me everything I really need to know - especially the patients I'm so close to that we are just about friends*". The sad reality is that, regardless of how supportive we perceive ourselves as being, our patients see us as the one's who can provide the care they desperately need if they can get us to do it.

A typical clinical example concerns phantom pain. The literature from before the 1980s indicates that between 0.5 and 1 percent of amputees have phantom pain. However, this was because the amputees did not volunteer the fact that they had the problem. When a physician did do a study in which he asked the patients if they had phantom pain, only a few percent admitted to it. Our studies showed conclusively that patients did not tell their physicians that they had phantom pain because they were afraid that the physicians would think they were insane and would not take them seriously when they reported problems in their stumps which require immediate attention or months of agony result (Sherman, 1997). These surveys were clearly being conducted by a group not related to the health care providers and was obviously anonymous.

In the clinical treatment environment, your subjects are your patients or your colleagues' patients. The patients came to the hospital for treatment - not to participate in a study. They are not going to take any chances with their care by telling you something you don't want to hear. They will do everything they can to keep you as happy with them as possible. This perceived need has ruined many studies. For example, very compliant post-surgical patients frequently won't tell the physician that they need more analgesia because they don't want to bother the doctor - who might get annoyed with their "whining" and not attend to their surgical problem as closely as they might otherwise. If the patient is participating in a study of analgesic potency, even sugar pills seem to work just fine. The placebo effect of sugar is overrated because of this and many patients have been left in agony.

Thus, a neutral team has to be brought in to do the asking - and the information needs to be gathered in as innocuous, anonymous way as is possible. This necessity extends to any outcome measures including taking blood pressures, measuring joint angles, etc.

4. If you don't ask the right question, you won't get the right answer: Continuing the example of misunderstanding phantom pain - Some years ago we (Sherman et al, 1980) sent surveys to virtually all physicians who were members of the American Pain Society and at medical centers treating amputees. Each was instructed to report the failed treatments for phantom pain found in their patients' records and to report the treatment they found successful along with their length of follow-up. Virtually all of the respondents felt that their treatment worked with nearly all of their patients - but virtually none did follow-ups. They thought their patients got better because they did not return for further treatment. As we all know, patients vote with their feet - if they don't get better, they usually don't come back if they have a choice - and they usually don't tell the practitioner. Further research showed that virtually none of the treatments were worth anything at all (Sherman 1996). Thus, these hundreds of physicians had fallen into the mistakes of (1) not following their patients long enough to find out if their

treatments were successful and (2) not testing their treatments using appropriate clinical studies.

5. Assumptions about prior work: Your study has to make many assumptions based on the accuracy of previous work. You can't check every study that has led up to your data gathering techniques. Unfortunately, Murphy's Law is very active in clinical research settings. Faked results and results based on populations very different than your patients are to be expected. Thus, the following are **likely** to happen:

- (a) You will miss at least one confounding variable until your study is completed.
- (b) Two of the basic studies needed to support your hypothesis and explain your results have not been done.
- (c) One of the basic studies demonstrating the validity of your data gathering and evaluation technique was faked.
- (d) Your norms are based on ten year olds tested in Mongolia during 1925 .

C. Creativity: Most clinical research can be divided into two orders of creativity - the cutting edge and the confirmatory stage. Confirming an earlier finding requires relatively boring, meticulous, lengthy studies with lots of control groups and details to clear up. The real adventure is at the cutting edge. You have an idea and want to test it carefully enough to find out if it has merit. But you are entering new territory. Perhaps nobody has really tried to answer this particular question before. There may not be well established ways to demonstrate the effect you are looking for. If you believe any of the above, you should be getting the hint that gathering good data can be difficult in the real world. Most of the problems you are likely to be interested in studying have been around for a while. You probably wouldn't be trying to study the problem at hand if someone else had already answered it to your satisfaction. Since other people must have looked at the problem, the solution must not be simple to reach or somebody probably would have answered it. In many cases, this is because the problem can not be approached directly. This means that you are going to have to really stretch your and your team's imagination to come up with a method of answering the question in a way that produces satisfactorily solid data.

For example, let's say you are a line Navy officer and your Admiral orders you to find out whether sea serpents are real. You are given a small team and a short deadline - but few resources to either question witnesses or go find a serpent for yourself. This really happened (Heuvelmans, 1968) and, as you might imagine, the Navy takes a dim view of young officers who can't come up with solid answers so the team felt under pressure to come up with something good. With little time and money to throw at the problem and no trustworthy experts to hire, they had to stretch their imaginations. What they did was one of the most brilliant pieces of simple detective work I know of. They started with the assumptions that (1) some but not most sailors actually aren't crazed drunks who simply invented stories, (2) most sailors didn't care who believed them or not so would just say what they saw, and (3) that if sea serpents were just figments of the imagination, the descriptions would be similar among people of similar cultures because each culture has an idea of what a typical sea serpent looks like (this goes for the

stigmata of Christ and many similar phenomena). They then looked up every report of sea serpent sightings which included both approximate location and description. They included every report they could find regardless of how many hundreds of years ago it had been made or the supposed veracity of the reporter. Everyone compiles lists of sightings of odd phenomena but few people find a meaningful way to organize them. This team did something novel. They plotted every report on a set of highly detailed navigational and ecological charts covering the entire world. They found that sea serpents were clearly distributed by ecological niche rather than specific part of the world or culture of the reporter. If they had used general maps, they would have missed the crucial information.

D. Techniques for hardening data:

1. Know your subjects and the disorder you are studying very, very thoroughly so you are not working in a vacuum. You need to know what variations to expect in your disorder and the impact placebos and now discarded treatments have had on it in the past. You absolutely must know how your subjects are likely to respond to the kinds of questions you are going to ask. For example, when attempting to determine the impact of phantom pain on amputee's lives, we asked them a series of questions about how they treated and used medical facilities to care for mild headaches and for a moderately severe pain that didn't resolve on its own. These data permitted us to gauge our subject's answers against those of the general population so we could get an idea of how open they were being.

2. Don't put yourself in a position where your subjects are likely to lie to you or you can fool yourself. Let a neutral team do the evaluations.

3. Failure of subjects to comply with complex and lengthy demands ruins many studies because the data are not accurate (or even real) and subjects refuse to continue past the point where they perceive any direct benefit to themselves with the result that follow-up data can not be gathered. This is especially true of complex logs that require multiple entries every day for months. There is considerable evidence that most patients stop making multiple entries in home logs after about a month and simply fill the log out once per day and then once per week.

4. Establish very clear response requirements for any information you elicit. For example, if you are asking patients to rate their pain on a scale of zero to ten, make sure they know what time period you want covered by the rating (that second, average of the day, etc.) and what the scale's end-points are (zero equals no pain and ten is so much pain you would faint if you had to endure it for one more second rather than ten is the most pain you can imagine which is not as clear an endpoint).

Define the variables you want people to give information on very carefully and clearly. For example, if you are asking for information about migraine headaches, make sure that you define the symptoms that you accept as a migraine so you don't get information about tension and other types of headaches. You would state explicitly what an aura is and give several examples. You would indicate that while actually vomiting with subsequent transient headache relief is an inclusion factor, nausea is not. You would explain what you mean by statements such as "describe what your pain feels like" so people don't tell you that they hurt. Give examples (e.g.

shocking, throbbing, tight, dull) but don't give them choices to check off or you may miss a crucial bit of information as their pain may not fit well into the choices you provide.

5. Always check the consistency of your measurements throughout the course of a study. Just about everything drifts with time.

6. Whenever two or more investigators are doing ratings, insure that they have practiced the ratings together until they give the same readings on both normal and a variety of abnormal situations. Continue to test inter-rater reliability throughout the course of the study in case the investigators' ratings gradually drift apart.

7. Practice your data gathering techniques until you have them down pat and the learning curve is over.

8. Use the most objective measures possible. One of them is likely to fail or have a flaw you do not pick up until it is too late in the study to replace it so try to use at least two measures. Be sure you test the measures against a gold standard as best as possible so you know they are working and what they relate to.

9. If you are using instruments, test their reliability, accuracy, and precision. They may not be up to your requirements. These concepts are discussed in the following chapter.

10. Do a pilot study and check to be sure your results are valid and reliable. Sometimes you have to add something that will cause a change in your test to insure that the test changes the way it should or to establish how the test responds. For example, if you are testing a method for picking up severity of an illness, check it with people who have established differences in severity.

Sample Study

Headaches are very subjective. What have the investigators done to make this data as firm as possible?

Chapter 19

Validity and reliability

(precision / accuracy of data and defensive data entry)

A. Setting subject expectations influences outcomes: Because people tend to know what they want when they design studies, extreme care has to be taken to ask the question in a neutral way and to insure that neutral parties objectively record the results with as little knowledge of what the participants went through as possible (e.g. which group they were in, what state in the study they are at, etc.). For example, early EEG Biofeedback studies in which subjects were trained to increase the predominance of alpha frequencies in their cortical spectra reported that their subjects felt more relaxed / serene after increased alpha. This was repeated several times until it turned out that the investigators were telling their subjects that they were likely to experience that feeling after increased alpha. Two studies showed this to be an artifact of expectation. In one, subjects were told that they would experience a change in emotional state but not what. Their subjects produced random answers. The other study told subjects that they would experience a change in emotional state but each subject was randomly told a state to expect. Subjects usually picked the state they were told to expect.

B. Precision / accuracy: Accuracy is how well the instrument/technique measures what it supposed to. Precision is how close to the same answer you get each time you measure the same thing. Along with reliability, precision and accuracy determine the internal validity of your study - the ability of the measurement technique to pick up the variable you are interested in. Obviously, no amount of precision or accuracy is relevant if the measurement technique is not related to the variable in the first place.

1. What is precise and accurate enough?: Your measurements must be sufficiently accurate and precise to provide the data you need or you are wasting your time. If you are doing a diet study, you would not ask people how much they weighed, you would weigh them. When you weighed them, you would use a scale which can reliably pick up differences of a quarter pound or so. You don't need precision of 0.1 ounce but it has to be better than plus or minus a pound or you won't be able to track changes adequately. Doesn't this seem ludicrously obvious? Why even bring it up, let alone harp on it by giving it a section? Because people are forever using measurement devices which are not accurate enough for what they need to do. The number of people who perform studies in which they have to do something such as measure angles to a tenth of a degree but use hand held instruments such as goniometers, which can not give repeatable measures within five degrees, it incredible. You must test your equipment to insure that it is reliable, precise, and accurate to the degree you require - and don't report the data to more degrees of accuracy than your instrument can measure. If you are measuring height to the nearest inch, don't report that the average height was 69.32914 inches. The best you could

extrapolate to would (generously) be one order of magnitude so report the average as 69.3.

2. Precision is not the same as accuracy: This is the idea that when you are shooting at a target you can get a very tight cluster of shots (the holes are close together) and you can repeatedly get that very tight cluster but you also repeatedly put that cluster in the top left edge of the target. You are precise but inaccurate. An example recently occurred in one of our studies which required a research associate to make a very complex measurement of over-pronation. There is no gold standard for the measurement and there is no machine which performs the test so she was trained by an expert to use a hand held goniometer (an angle measuring device) to make the required measurements. She had to measure thousands of people quickly while producing consistent, accurate measurements under difficult circumstances. We checked the reliability of her measurements by having her test a random group of the subjects a second time several days after the first test. The test-retest reliability of the associates' measurements was 0.92 (out of 1.0 - which is incredibly high for a clinical measurement done with hand held goniometers) with one of the eighteen retested subjects changing from meeting to not meeting the study's entrance criteria and the other changing the other way. Both were borderline at the first measurement. However, we also had the highly experienced expert come back after the study was in progress and measure a group of the subjects within moments of the research associate taking her measurements. The associate's measurements were consistently lower than the expert's by a ratio of 0.56 (standard deviation of 0.225). This was enough of a difference so that many people who should have participated in the study did not because they apparently did not overpronate sufficiently to meet the entrance criteria. Thus, her readings were consistent and reliable but not sufficiently accurate for the study to be performed properly. In retrospect, it is obvious that we should have figured out some way to calibrate her measurements so she remained consistent with the measurements she made when she was trained initially. For instance, we could have had both the associate and the expert make several initial measurements on over-pronating staff who were to be present throughout the course of the study and then have had the associate repeat those measurements every week or so. Unfortunately, we didn't realize that her measurements would drift once the study got underway.

3. Do you want to be as precise as possible? Sometimes you have to balance the intrusion required to get a highly precise measurement with the impact that intrusion has on other aspects of your study. You may wish to have people log their pain five times per day but there is no use asking for this level of precision if they are going to drop out after a week because they don't want to keep the log. Very occasionally it may be a good idea to gather somewhat less precise data than could be gathered for clinical reasons or to strengthen other parts of the study. For example, in a study on migraine headaches initiated by an aura, you may need to know the duration of each headache. However, if you have your subjects start a timer as soon as the headache comes on and stop it when the headache goes away, you may get very inaccurate data because your subjects will be attending to the headache. They may keep the timer going as long as any dregs of any head pain are present. The odds are that the actual headache you are studying will have gone away hours before many of the subjects stop the timer.

If the measurement system is quite intrusive, you risk engendering the Hawthorne effect in which subjects behave differently because you are watching them or that they are special participants under study. This is the Achilles's heel of many time-motion and kinesiology studies as the participants know they are being watched and do not behave the way they usually do.

C. Reliability:

1. Test-retest reliability: Your instrument must be sufficiently reliable to repeatedly give measures within the level of error you can accept. The need for test-retest reliability assessments is generally well accepted when evaluating the ability of a therapist to consistently make a measurement but tends to be skipped when a machine is making measurements. If you are weighing people, you can not assume that the scale will give you the same answer every time the same weight is put onto it. You need to test the scale yourself and find out what the expected variability is.

2. Blind ratings: Most outcome measures are not entirely objective. E.g. when reading an X-ray, the reader will look extra hard for a problem if it is known to probably be there. Thus, the rater needs to get a mix of normals and diseased subjects.

3. Interrater reliability: If several raters are working, they must agree on the criteria and rate several of the same subjects (with different intensities of the problem) so that their ratings will match. Having the group rate a disproportionate number of normals is a problem because raters frequently agree more on normals than on abnormals. Disagreement tends to be less when there are distinct yes/no types of judgements rather than the need to make qualitative (0 -10) ratings. It is also lower if the subjects are not as sick as the ratings all tend to be in a smaller range. Disagreement tends to be greater as the number of possible outcomes increases. Of course, the rate of disagreement increases as the number of raters increases. The crucial need for pre-establishment of mutually recognized criteria was demonstrated recently when a group of physicians attempted to demonstrate that trigger points could be reliably picked out on the upper bodies of a group of patients by having each physician see each patient in random order within a few minutes of each other. The physicians did not work together on several patients the first time the study was attempted - and it failed. The second time, they did reach mutually acceptable criteria and the study succeeded. They were usually able to identify trigger points in similar places on each subject.

Interrater reliability is usually measured by correlating the readings each rater gives with those of the others. The usual statistical methods for evaluating consistency among examiners are (a) linear or non-linear correlations, (b) the interclass correlation coefficient, (c) Cohen's kappa, and (d) Cronbach's alpha (Bordens and Abbott 1991; Turk and Melzack, 1992). The interclass correlation is for continuous measurements and the kappa is for categorical measures.

(a) Linear and non-linear correlations: For continuous measures, I tend to use a standard Pearson's correlation and for non-parametric rating scales I use a Spearman's correlation. This is not as sophisticated as the other tests but it gives me a very good idea of how close the raters come to each other. The coefficient of correlation gives you the amount of variability due to differences in ratings. If the correlation is 0.90, then 90% of the variability is explained by differences in the ratings so the raters have come very close to each other. The method has the strength that the level of statistical significance is calculable so you know the odds of how strong the relationship is. However, the technique has the weakness that

correlations can be very high without the raters actually having agreed exactly. They just have to be consistently close. For example, if observers one and two both give ratings of 1, 5, 9, and 15 the correlation will be perfect (a 1.0). However, if observer one gave those ratings and observer two gave ratings of 2, 6, 10, and 16, the correlation would still be a perfect 1.0 even though they are not the same. Thus, you need to do a χ^2 test or its non-parametric equivalent to make sure that the ratings were not significantly different in magnitude.

(b) The interclass correlation uses the variance error terms from an analysis of variance to sort out variation due to the subjects and the examiners. Some statistical analysis programs include this statistic, but, if not, you can use the following formula to compute it:

$$\text{Interclass correlation} = \frac{(\text{Mean square error for subjects} - \text{overall mean square error})}{(\text{mean square error for subjects} + (\text{number of examiners} - 1) \times \text{overall error} + ((\text{number of examiners} \times (\text{mean square error for examiners} - \text{overall mean square error}))/\text{number of subjects})}$$

This formula is essentially a ratio of the variances so it represents the amount of variability explained by the subjects relative to that explained by the examiners. The formula produces a decimal which is the correlation coefficient. If the coefficient is 0.90, then 90% of the variation is due to the subjects. The remaining 10% is due to the examiners and measurement error.

(c) The kappa coefficient is only for categorical data and compares the observed agreement between two examiners to the amount of agreement which would be expected by chance alone. Many statistical packages calculate this statistic. If you need to do it by hand, the formula is:

$$\text{kappa} = \frac{(\text{Pa} - \text{Pc})}{(1 - \text{Pc})} \text{ where}$$

Pa = actual proportion of agreement
Pc = chance proportion of agreement

The calculations, if not the logic, are easier to follow with an example. In the following example, two raters (one and two) rated two variables (a and b). However, in the formula, any number of raters and any number of variables rated can be used. For more than two raters or variables, in any instance where adding or multiplying is called for, just use all the appropriate data.

The actual proportion of agreement is the number of agreements over the total number of observations. For example, if there were 40 observations and the two raters agreed on some combination of 38 of them, then Pa (actual proportion) = 38/40 = 0.95.

The proportion of agreements expected by chance (Pc) is more complex to figure out. It is the sum of the products of the number of times the two raters used variable a and the number of times they used variable b all divided by the square of the number of observations. Let's say that rater one used variable a 34 times and rater two used it 32 times. The product is 1,088. If rater one used variable b six times and rater two used it 8 times, the product is 48. Now add 1,088 to 48 to get 1,136. In the example, there are 40 observations so $40^2 = 1,600$. Now divide 1,136 by 1,600 which is 0.71 which is Pc.

So the $P_c = 0.71$ and $P_a = 0.95$. Filling in the above formula, $k = (0.95 - 0.71)/(1 - 0.71) = 0.83$.

The score could be anywhere between -1 and +1 but negative scores indicate agreement levels below that expected by chance alone. The higher the kappa score, the greater the level of agreement. Anything over 0.8 is considered excellent. The two raters in our example weren't particularly good. However, many studies have been published with kappas hovering around 0.3 and were considered marginally acceptable.

4. A few ways to increase precision:

- (a) Standardize your measurements using an SOP (standard operating procedure).
- (b) Conduct structured interviews instead of open interviews:
- (c) Train all observers the same way, check them against each other, and recheck them all against the same standard periodically throughout the course of the study.
- (d) Refine the test instrument through pilot studies and use in as real a set of conditions as possible.
- (e) Automate everything you can - and still check to make sure the automated device is working correctly. This is especially true for having subjects fill out questionnaires when they are in your clinic. A computer presented questionnaire can guide the subject through the instrument so questions can't be skipped and so answers outside the accepted range can be brought to the subject's attention immediately.
- (f) Repeat the measurements often enough at each measuring session to establish variability and reduce the likelihood of recording a fluke bad measurement.

D. Calibration: Everything breaks down and measurements made by both humans and instruments tend to drift with time so you need to check them on a regular basis. I recently had the embarrassing experience of being the person responsible for insuring that an instrument was working and calibrated and then went on a two week trip without getting a replacement to perform the task. The machine in question had worked for over a year without a hitch so I didn't bother. Of course, the gizmo got out of wack while I was away and a whole series of subjects were exposed to half the dose they were supposed to get. We discovered the problem and added an unanticipated dose-response section to the study which gave us important data - but it was still embarrassing (especially since I had to admit the lapse in the subsequent article).

E. Data entry - the big error few people catch that randomizes the results!!! Manual data entry is plagued with inaccuracies. For example, our research team had 25 amputees keep daily logs of their phantom pain intensity, the weather, stress and other factors which we thought might effect their pain. The subjects kept their logs for six months or one year. Each log sheet covered two weeks and contained 378 cells of information. A minimally skilled technician was

paid minimum wage to transcribe the data from the log sheets to a computer data base. The technician was instructed to check all of the entries after each log was entered. He was a steady worker who certainly appeared to be doing a very careful job. There was a pause between data entry and analysis during which the technician left the area. The student (Amie Urton of Evergreen State College) who was to perform the data analysis began by doing a through check of the accuracy with which the data had been transcribed. She found several types of transcription errors including entering an entire row or column in the wrong spot and entry of incorrect numbers. The error rate for misentering a number - either miskeying or misreading the number and then entering the wrong number in the right place (not counting incorrect placement of columns or rows) was 2.4% in one of the two week logs. It took Amie and a colleague working together two hours to check and correct one patient's six month record. This error rate is typical of what we find when uninterested, uninvolved people enter boring data from complex formats.

If you must enter data manually, the most accurate and best way to avoid mistakes during manual data entry is to use the buddy system. One person reads the data from surveys, etc. while a second person enters it. The second person reads off what they are entering as they enter it so the first knows that it was done accurately.

Many data entry programs can be set so they give a warning when you make an entry outside the preselected choices or limits. This is well worth doing because it can catch numerous typographical errors.

Data cleaning is a special stage in the data entry process. In this stage, you look for logical inconsistencies in the data such as a 15 year old having broken his leg 36 years ago. When developing algorithms for how you are going to clean your data of obvious typos and pick up errors made by the respondents, you need to be very careful not to make your acceptable limits too large or include too many differing subjects in each pass through the data. It is better to pick up and then OK many good bits of data rather than miss a few bad ones. For example, if you are studying anorexia and mix the weights of males and females together, you could miss an outlier female weight because it blends well with normal male weights.

F. Data Security - Storing data so it can't be changed or lost:

1. Preventing loss: Murphy's law does not appear to strike randomly. It strikes best when its effects have been enhanced by our own carelessness. Virtually all research data are now stored on computers at some stage in their reduction. Even when a hand written set of original raw data exists, it is stunningly difficult to reenter a major data set into a computer by hand. The reality is that losses really occur because of computer disk crashes, back-up disks failing, hand written raw data being lost or ruined, etc. I suspect that everyone who does research for any length of time has had unfortunate events happen to their data. I have lost hand written raw data files when they were thrown out accidentally and their loss was not discovered until too late. I have also had a hard disk crash and then discovered that one of the back-up disks had something wrong with it. Thus, take data back-up seriously. Photocopy your raw data as you gather it and keep the copies at a distant location such as home. Always back up everything you put on a hard

disk. I back up my data every day it is changed and make a second back up which I keep at home not less than once per week. The number of times I have had to bring in the home files says a great deal about how fragile our computer based storage routines really are. It doesn't matter what type of magnetic media you back up your data on as long as you have at least two machines which can read it. I know one person who was backing up data on a tape drive which went down with the computer. It turned out that nobody nearby had a tape drive which could read the tape.

2. Preventing unwarranted alterations: It is very tempting to make "minor" changes in data to eliminate "obvious" outliers, misreadings, etc. Investigators who enter their own data are going to cheat if they wish to and there is nothing that can be done about it unless an audit of patient records containing independent data is carried out by the investigator's institution, NIH or etc. The more usual situation is that changes are made either on purpose or inadvertently by technicians who are responsible for entering the data. There is little that can be done about the situation where someone writes down the observation incorrectly in the first place since there is no record of the actual number anywhere. The best defense is to convince everyone participating that whatever happens, happens and that the results are just as good either way. All hand written data should be written into data books with numbered pages which are signed by the person entering the raw data. Any changes are initialed by the person making them and each participant is responsible for keeping their data book secure. As noted above, the data should always be entered by two people but not only for accuracy. It is less likely that two people will cheat than one. Just as with data entry, random checks comparing the raw data sheets with the computer data base are crucial to pick up any differences which are present for whatever reason. The data entry program should require a secure code to enter and a back-up copy should always be kept secured out of the data gathering and entry area. This is important not only for security but in case of fire or other misadventure.

G. The art and science of data management - the professional data manager or how to avoid or handle reams of data the FDA and others want:

This is essentially a warning. If you get involved in a large study which is going to involve following hundreds of patients over months of repeated measurements made by numerous different clinics, a typical clinician is not going to do the data management job adequately and still do the clinical job. The FDA and other governmental groups require astonishing amounts of detail - and they send inspectors with police powers out to make sure you are gathering it. Thus, if your large study is part of an effort to get a treatment or device approved by the FDA you are going to have mountains of information which must be organized in a very precise way. The usual reaction to this kind of study is to hire (or otherwise obtain) a nurse to do the job. They can do an excellent job gathering the clinical data but don't have a clue as to how to organize, enter, and store it. The same happens when clinical support staff are pressed into the role of data manager. Before harming your project and devastating the nurse or support staff this way, send people for training. This costs a few weeks of people's lives and a few thousand dollars but saves projects. If you have a very large project or several medium projects, you need someone who is actually trained and experienced in how to do data management. Let them manage your data and leave the gathering of clinical data to your nursing and support staff.

H. Bias - Vital criteria for avoiding bias in performing outcome studies: The perils of not obeying these criteria were exemplified for natural history studies of nonunion of the scaphoid in an article by Leslie Kerluke and Steven McCabe (1993) in the the Journal of Hand Surgery. The following list is modified and extended from that article.

1. Inception cohort: Start your study with everybody in your referral base who has the disorder. If you start with people who develop a secondary problem (such as people who are still symptomatic after six months, who develop arthritis, etc.) you have three potential biases:

(a) You miss the patients who do not develop any problems and do just fine with the initial treatment (e.g. they never developed arthritis after the Rx.).

(b) If you start a study well after the treatments have been given, you may introduce a selection bias because you can't enter those who died (from the treatment, disorder, etc.) or have disappeared for other reasons. The sickest people may not have survived long enough for you to determine whether the treatment was effective.

(c) If you do a cross-sectional recruitment design (e.g. recruit from people who happen to be in your office over a two month period regardless of when they had the surgery), you have a greater chance of seeing people with many problems as they are more likely to be in your office.

2. Representative referral pattern: If you are in a tertiary care center or a VA, you do not see typical patients but, rather, ones likely to be more challenging due to age, other problems, etc. so your results will not be as good as those of a group working with relatively healthy patients.

3. Complete follow-up vs. missing the most successful outcomes because they disappear and do not ask for further care.

4. Blind ratings and Interrater reliability as discussed above.

5. Objective criteria: People see what they want and expect to see. The more objective the criteria, the better the chances of an accurate result.

6. Control for factors likely to affect the outcome - e.g. age.

7. Bias in collecting the data if the investigator collects it as discussed above.

I. Choosing outcome measures - they have to make sense: What you decide to measure determines the external validity (how closely the variable is related to the problem) of your study. Sometimes the actual problem you want to measure can't be measured directly. For example, pain can not be measured directly. If you want to know about perceived pain intensity,

you could ask the person to rate their pain but if you want to know about the effect pain has on the person's life style, you may want to do more than just ask. You may wish to interview family and colleagues to find out if changes in the patient's ratings of pain correlate with changes in family interactions and work. Numerous studies have shown that asking patients how physically active they are just after surgery does not correlate well with data from nurses or pedometers. The nurses ratings do not correlate well with the pedometer data either. Thus, you must check on the validity of your outcome measure to be sure it matches what you actually want to know.

The nightmare begins when you don't realize that your have chose an outcome measure which changes disproportionately to changes in the problem you are trying to assess.

Chapter 20

Survey, test, and questionnaire design

A. Overview of data gathering methodologies: When clinicians want to know how their treatments are going (effectiveness, side effects, etc.), they usually ask their patients in one way or another. It has been well established that patients rarely tell their health care providers the truth about their progress or even about what is actually wrong with them. For example, as recently as a decade ago, only about 1/2 of one percent of amputees would tell their health care providers that they had phantom pain without being asked and only about five percent would admit to it when asked directly because they were afraid that their health care providers would think they were insane (Sherman, 1997). Patients usually do not tell their health care providers when a treatment has not worked, they just go to a different practitioner. Unfortunately, many health care providers are left with the impression that their therapies are effective when, in fact, they are not (Sherman et al 1980). This is one of the main reasons that clinicians can not gather subjective outcome data from their own patients. A separate team has to be involved in the assessment system. Thus, clinicians can not find out how their treatments are working or about underlying problems by simply asking the patient. It is also very difficult to get together with a large group of patients seen over a period of time (perhaps even years) when sufficient time is available to get in depth information.

These limitations on patient contact result in the need to use some sort of survey technique to get the required information. This requirement becomes even more obvious when population baseline information has to be gotten from people similar to the patients in question. This is especially common when the natural history of waxes and wains in a disorder have to be established.

The five most common ways a neutral assessment team can survey people without having an involved clinician actually ask them are:

1. Personal interviews
2. Group interviews
3. Phone interviews
4. Questionnaires given out personally
5. Mail-response questionnaires

Each of these ways of finding out what is going on have their own strengths and weaknesses.

B. Personal interviews: The advantages of personal interviews include the ability to establish rapport with the subject to increase the depth of variety of answers. New information can be

discovered which the investigators were unaware of so couldn't ask about in a formal survey. An other advantage is that in the clinical setting nearly everyone agrees to participate in a personal interview when it is conducted while the potential participant is waiting for something such as pharmaceuticals or a medical appointment. The strongest disadvantage is that the interview can not be truly anonymous and the subject has to be willing to tell you information which the subject believes might be disparaging to him/her in your eyes. For example, if the subject is an amputee who believes that you would think he or she is insane if phantom pain is reported, you aren't likely to hear about it.

Interviews can be highly structured, unstructured, or some combination of the two approaches. Most interviewers in the clinical environment work from a check list so they can be sure they ask at least the crucial questions but leave lots of space for encouraging comments. Information gleaned from interviews is frequently used to establish and check questionnaires.

Interviewers have to be non-threatening, pleasant, and encouraging. They need to present an air of interest and patience or the subjects will not cooperate. There is nothing worse than an interviewer who acts as though he/she already knows everything, seems to be a **haughty** clinician the patient may meet the next time they need care, and is in a hurry. To be blunt, in my experience, the optimal interviewer in the clinical setting seems to be a young, pleasant, easygoing (non-threatening) woman who gives the appearance of interest and sympathy while not being too knowledgeable. This is illustrated in everyday clinical life by the vast difference in what patients share with technicians versus what they share with the □doctor.□ For example, the discovery that migraines are prevented by exposure to pulsing electromagnetic fields (see sample study) was made by a research associate in the course of apparently informal conversation with subjects (read this as habitual establishment of good rapport) rather than by the project director who asked the patients every week whether anything at all in their bodies had changed.

Individual interviews are the most time consuming way to get information and, on a per-subject basis, are the most expensive. They are not useful when information has to be gleaned from masses of subjects so they are restricted to situations in which depth and flexibility are required.

C. Group interviews: Group interviews and "sensing" sessions are becoming more common in the clinical setting. Their strength is that people together may well feed off of each other and come up with more ideas than they would alone. The subjects are also more confident when in a group than when alone with an interviewer so are more likely to come out with information they wouldn't give away otherwise. This is especially true once one subject breaks the ice and says something that the rest of the subjects know but which nobody ever tells the "doctor". I have found this milieu excellent for getting amputees to talk about problems they are having with their prostheses because they "egg" each other on with examples and one person's story jogs other people's memories.

Selection of participants for group sessions is an art in itself because the correct mix of subjects have to be gathered in one place at one time. Availability of subjects can easily skew the responses away from people who feel they are too busy or uninterested. The major weakness is that one or two dominant personalities may take over the group so more reticent subjects may not

be willing to present different or conflicting information. Thus, crucial data can be lost and a false impression of homogeneity can be created. It takes a very skillful facilitator to avoid these pitfalls. Another weakness is that the semi-structured interview technique can not be used as it is very difficult to direct the conversation. Only a limited number of topics can be covered so this technique is optimal for exploration but of little value for gathering lots of detailed information in a brief period of time. This is a fairly expensive way to get a moderate amount of data from a few people.

D. Phone interviews: This is a very tricky technique rife with pitfalls. Phone interviews are usually conducted either to (1) establish the rate and characteristics of a problem in a population, (2) identify characteristics of a problem from a group identified as having the problem, or (3) follow-up with specific people who have been treated for some specific problem.

1. Population interviews: These are performed to establish the rate and characteristics of a problem in a defined population. A typical phone survey might be conducted by an HMO to get an idea of how many people at a large corporation joining the HMO have low back pain they normally receive treatment for. Phone surveys also attempt to determine the incidence and impact of a disease in a general population such as headaches in all of New York City.

a. When using this technique to elicit information from a large population, such as everyone in New York, problems inherent in this technique include how many people to call, how to select phone numbers, who to talk to when someone answers to phone, and how to get the people to talk with you. As recipients of your calls will assume that you know who you called, they will respond as though this is not an anonymous situation so their answers will be skewed. Although it is true that nearly every household in New York City has a phone, the number of phone lines is not randomly distributed (poor people have less) and the very poorest people don't have a phone at all because they don't have a place to live either. Usually a random number dialing program is set to dial a specific number of phones in each neighborhood of the city being surveyed so all economic strata are phoned. Calls are made at a variety of times throughout the day because different age and socioeconomic strata are available to answer at different times of the day. Once someone does answer, it is necessary to get the type of person you want to talk with onto the phone rather than simply accepting answers from anyone who picks up. This cuts down the number of answers but increases quality and specificity. For example, if you are trying to determine how many people in New York get headaches and their impact on day of work lost, it is of little use speaking with the six year old who answers. If many questions are to be asked, it is usually best to offer to call back at a convenient time. Once you have completed your survey, you need to check the distribution of the people who answered because, although you may have dialed a proportionate share of neighborhoods, you may have answers from a disproportionate share of one or another which could lead to a disproportionate number of answers from people of one socioeconomic group. Determining how many calls to make is discussed later in the chapter in the section on sample size.

b. When using this technique to survey an organization, similar problems apply to those identified above for surveying an entire city or country. The odds of people speaking with you and giving you valuable information are increased by identifying your call with the

organization you are surveying and explaining why you need the information. The respondent should realize that providing the information might have at least some long term benefit to themselves so their motivation to cooperate is likely to be higher than that of people called at random. On the other hand, call recipients are likely to be far more threatened by the call than recipients from the general population because they may fear that their answers could affect their jobs or health insurance. The odds of anyone actually believing that you are not recording who you spoke with are about zero. This must be taken into consideration when interpreting the data.

2. Surveys of a specific group identified as having the problem of interest: A typical example would be to survey everyone identified by a hospital's data base as having been treated for migraines. This is a powerful way to find out about success of treatment, variety of treatments used, and characteristics of the disorder. The weaknesses are (a) that you have to be able to get a representative sample of people on the phone and (b) computer data bases tend to be inaccurate so many of your calls will be to people who don't actually have the problem. I have used this technique very successfully for exactly this purpose with considerable success.

3. Surveys of patients treated for a specific problem: This technique is used to find out how successful a treatment has been. It tends to be used when only a small group has been treated so a maximum response rate is critical to getting useful information. It is especially valuable for asking quality of life questions because patients have the flexibility to tell you information you might not think to ask. Of considerable importance, these surveys can be done longitudinally. The first call can be made before treatment is initiated with subsequent calls being pre-arranged to follow just after treatment termination and at logical follow-up periods. Once a patient is primed for a series of calls, the odds of their cooperating are very high. They are far more likely to answer a phone question than a mail response survey under these conditions.

E. Mail response questionnaires: Mail response questionnaires are among the most commonly used tools to gather "clinical" data. Unfortunately, they are also the most difficult to use properly and rarely produce valid results. Questionnaire methodology is a complex art and science. There are entire books on questionnaires and most study design books have at least a chapter on them.

The most common use of mail response questionnaires in a clinical setting is the treatment "follow-up" format. The health care provider usually wants to know how well the treatment works and how it relates to quality of life. There is little use performing an expensive, difficult, risky hip replacement if the recipients don't experience significantly (to them) less pain and more mobility for a reasonable amount of time.

Among the critical problems with mail response questionnaires are (a) low response rates, (b) skews in who responds so that the responses do not represent the population being surveyed, (c) failure to validate the questions (do the subjects understand what you are asking?), (d) poor design of the actual page, and (e) failure to design the questionnaire so it can be analyzed in a way that will provide the information actually required.

1. Enhancing the response rate: Most people consider themselves fortunate if half of anonymous surveys are returned properly filled out even when they are sent to a specific target audience which should be interested in the clinical problem being asked about. People who had a

hip replaced should intuitively recognize the importance of a questionnaire arriving a year after the surgery which is asking about changes in their quality of life relative to the surgery. If the relationship is made clear in a brief PERSONALIZED cover letter from the surgeon who did the replacement (presumably an important, relevant person to the recipient), and the survey is anonymous, brief as well as to the point (no extraneous questions), and clear, there should be a reasonable response rate. A poorly photocopied cover letter without the patient's name typed on it which is stapled to a poorly copied survey asking obscure questions not obviously related to hip problems which is so poorly designed it can't be answered or contains inflammatory or threatening wording is very likely to wind up in the trash can along with surveys which are more than a page or two long or which insist on patient identification without clearly indicating why anonymity has to be broken.

2. Testing and designing the survey: Clinicians and patients tend not to use the same terms and have the same concepts for any particular problem. Even if the clinician asking the questions has had the problem being studied, he/she has only his/her personal (□n□ of 1) experience to draw from. In order to make up a survey which will be intelligible to average patients and ask nearly all the right questions, a six step approach is needed.

First, the clinical team makes up the first draft of the questions by using their own experience and a solid literature search of both the clinical problem being investigated and other surveys covering similar problems.

Second, the questions are shown to other clinicians familiar with the problem so they can comment on potential answer choices and information left out. It is important to work on the vocabulary, concepts and layout of the questions before showing the survey to other clinicians so they won't concentrate unduly on structural matters but, rather, will be able to focus on medical information.

Third, at least one "focus" group of patients having the problem needs to be held to get their ideas of what is important to have in the questionnaire and to comment on the questions asked so far. It may be surprising what they feel is important and how different their understanding of the problem and the treatment is than the clinician's understanding. The focus group should also be asked what the most effective cover letter would say and who it should be from.

Fourth, the survey is put into a distributable format and a draft cover letter is made up. Questions have to be very easy to answer. If subjects have to try to figure out where their answers go they will either dump the questionnaire or do a poor job. See Figure 3 for examples of poor and better styles of question layout. The questions have to be very clear and specific so people know exactly what you are asking. Table 8 contains examples of poor and better question construction. Numerous surveys are longer than most people will put up with answering and ask questions which appear to be useful upon first glance but are either irrelevant or give such complex information that their results are never used because thousands of people would have to answer for sufficient responses at each choice provided to permit analysis. Most people throw away surveys which are more than a page long, are poorly set out (look amateurish), are poorly reproduced (lousy photocopy job), look like a computer test,

and - most importantly - do not have a brief cover letter from someone significant to them saying why their answers are important to their own future care.

Fifth, test your survey on about ten patients who have the problem being surveyed. They need to be as similar as possible to the people to whom you are going to send the questionnaire and they should not have been in the initial focus group because they will have an idea of what you are trying to do. Have each patient fill out the survey individually. Then go through the survey with each patient alone. Ask what they thought each question meant and what their answer meant. Ask for better choices for each question and for questions left out. This information is used to modify the survey one more time.

Sixth, mail the survey out to a trial group of at least one hundred patients from the group you are actually surveying to establish response rates, variability in question responses, and problems with the survey mailing and processing system. Response rate and variability information are used to determine how many surveys you actually have to send out to get a meaningful answer to your question. Once you incorporate this information, you are ready to send out the full survey.

A crucial note about survey questions: **DON'T ASK ANYTHING YOU DO NOT NEED TO KNOW** to get the information you actually need. The shot-gun approach makes long surveys and people don't answer long surveys with obviously irrelevant, personal questions. If you don't need income, don't ask. People don't like this. Remember to design your questions around the actual information you are trying to get and leave the rest out. You can always send a subsequent questionnaire to a subgroup if you need to.

3. Establishing and checking validity and reliability of the questions:

a. Reliability is the idea of how likely it is that respondents will answer honestly and accurately. This is best tested when you give the pilot survey to the group you will individually interview. For those surveys given twice, you can see if people answer the same question twice the same way. You can also ask the same question two different ways to see if you get the same answer. This type of check for internal consistency (sometimes called split half reliability) is very difficult to do without insulting the respondent. As soon as people see that questions are being repeated, they realize that you don't trust them and tend to purposely answer differently or dump the survey. The best way to approach the problem is to find two variables which you know will co-vary with each other very highly but which the subjects may not realize do so and ask about them both. A very effective way to check reliability when surveying patients whose records are available is to ask something for which you know the record to be accurate and see if the patient gives a similar answer.

b. Content validity is the measure of how closely related the answers to your questions are to what is actually happening to the patient. This is established through the sensing groups and interviews discussed above.

c. Criterion validity is the correlation between the answers to your questions and some "gold" standard. For example, ask if the patient has seen a doctor at your facility for treatment of a cold within the last year and then check the medical record.

Figure 3

Formatting the Questionnaire

A. Can't line up the answer categories with the questions:

Poor format

3. Rate the following on a scale of 1 - 8 where 1 is the least and 8 is the most:

	1 - 8
a. slslkdkdkljfkjfg g	00000000
b. rieief ffj k klkdfk	00000000
c. hlkljluoto jrj jt t t ty	00000000
d. qw ri nynntrn ntnt	00000000
e. iotmfm y mrmr yy	00000000
f. dkei di hg i i fkkgk	00000000
g. az x cnmddr ttyy t tt	00000000
h. sk gk yyyu ioop u ik	00000000
i. dkdk eir df fff	00000000

The above sample is ubiquitous in questionnaires. Nobody can tell which line of bubbles relates to which question and it is nearly impossible to tell which bubble in the row corresponds to which number. The bubbles are placed so far to the right because the antique scanning systems used to require it.

More answerable format

2. Rate the following on a scale of 1 - 8 where 1 is the least and 8 is the most (circle the number which corresponds to your rating):

a. slslkdkdkljfkjfg g	1	2	3	4	5	6	7	8
b. rieief ffj k klkdfk	1	2	3	4	5	6	7	8
c. hlkljluoto jrj jt t t ty	1	2	3	4	5	6	7	8
d. qw ri nynntrn ntnt	1	2	3	4	5	6	7	8
e. iotmfm y mrmr yy	1	2	3	4	5	6	7	8
f. dkei di hg i i fkkgk	1	2	3	4	5	6	7	8
f. az x cnmddr ttyy t tt	1	2	3	4	5	6	7	8
h. sk gk yyyu ioop u ik	1	2	3	4	5	6	7	8
i. dkdk eir df fff	1	2	3	4	5	6	7	8

Figure continued on next page

Figure 3 Continued

B. Questions easy to miss on the page:

Poor format

1. sex: ____ age: ____ DOB: ____
marital status: ____ race: ____ yrs of school: ____

More answerable format

1. About you:
a. Which sex are you? male ____ female ____
b. Your age in years: ____ (years)
c. Your date of birth: day ____ month ____ year ____
d. Your **current** marital status: married ____ single ____ other ____
e. Your ethnic background / race: white ____ Hispanic ____ other ____
f. Check all levels of school you completed: elementary school ____
middle school ____
some high school ____
high school ____
more ____

C. Can't easily tell which area to check off goes with which answer:

Poor format

2. I live with my mother: O yes O no O not sure

More answerable format

2. I live with my mother: yes O no O not sure O

D. Poor reproduction:

This says to throw the survey away because you didn't care enough about it to reproduce it nicely. If it doesn't look professional, it doesn't seem to be important enough to answer.

Table 8

Wording of questions

These questions would be from a post-treatment questionnaire following treatment for peripheral vascular insufficiency.

Poor wording	Problems	Better Wording
<p>How much further can you walk since the end of treatment? _____</p>	<ol style="list-style-type: none"> 1. People are terrible about judging distance. They are better at noting passage of time. 2. No idea of why they stop - run out of breath or pain in legs. 3. The question is not neutral - it assumes benefit. 4. Immediate post-treatment period is mixed with time when treatment had a chance to take effect. 5. The measurement scale the subject is to use (blocks, miles, etc.) is not indicated after the question. 6. There should have been a pre-treatment survey with questions worded about the same way so pre-post data can be compared. 	<p>During the last 5 days, how many minutes could you usually walk before your legs become painful? ____ (minutes)</p>
<p>Rate your pain on a scale of 1 - 5: _____</p>	<ol style="list-style-type: none"> 1. Not specific enough about the circumstances of the pain or when (after walking, all of life, etc.). 2. The scale is non-validated so can not be compared with known indices and may not work. 3. The scale has no anchor points so patients know how to use it (e.g. 0 = no pain). People are used to zero being none. 	<p>What is the worst your legs have ever hurt after walking during the most recent five days? (Use a scale of 0 - 10 where zero is no pain and ten is so much pain you would faint if you had to bear it for one more second.) _____ (1-10)</p>
<p>Your race: White__ Black__ Hispanic __</p>	<p>What if you are not one of these choices? Always leave a category for <input type="checkbox"/>other<input type="checkbox"/> or have an open ended question so people can tell you what is happening.</p>	<p>Your ethnic background White__ Black__ Hispanic__ Other (what) _____</p>

<p>Rate the quality of the food in our hospital: very unhappy___ not happy ___ somewhat happy___ very happy___</p>	<p>1. The answers don't match the question. This tends to happen when you try to use the same scale for many questions. This tends to confuse the respondents. No only is the validity of the answer in doubt, it lowers your credibility - which, in turn, makes it less likely they will complete the survey.</p> <p>2. The scale isn't balanced. There are more poor choices than good choices around the center value. In the example to the right, you would not add an <input type="checkbox"/>awful<input type="checkbox"/> category without adding an <input type="checkbox"/>excellent<input type="checkbox"/> one.</p>	<p>Rate the quality of the food in our hospital: did not eat at hospit___ poor___ fair___ good___</p>
<p>Rate your therapy on a scale of 1 - 5:___</p>	<p>Poor wording and too nonspecific. What did you want to know?</p>	<p>Rate how much activity you did during physical therapy during the first week after your surgery on a scale of 0 - 5 where.....: ___ (0-5)</p>
<p>Rate the intensity of crepitus in your diarthrodial joint on a scale of 1 - 5:</p>	<p>non-medical folk (and lots of medical folk) don't have this vocabulary!</p>	<p>Rate how much grinding you feel when you bend your knee as you stand up from a sitting position on a scale of 0 - 4 where zero is no grinding and four is so much it locks your knee before you can straighten it out: ___ (0-4)</p>
<p>Were the nurses on your ward courteous and efficient? ___</p>	<p>What if they were very efficient but terribly impolite?</p>	<p>About your nurses: a. Were they usually polite to you? ___yes ___no b. Did they help you when you needed it? ___ yes ___no c. They provided the information I needed:</p>

		<input type="checkbox"/> always <input type="checkbox"/> sometimes <input type="checkbox"/> never
--	--	---

F. Questionnaires given out personally: Giving out questionnaires in person avoids many of the pitfalls inherent in mailing them out. You know exactly how many surveys got to people and you can be reasonably certain that they went to your target audience. For example, I frequently have surveys given out to people waiting in our pharmacy to pick up prescriptions. Asking patients to fill out a brief questionnaire about some aspect of their health or medical care is a very different proposition than giving out surveys to disinterested, hurried shoppers at a mall. Just about everyone is willing to fill the surveys out (especially as they have nothing to do while waiting and are thinking about their health problems). They are aware that the person handing out the survey outside the pharmacy can't tell whether they actually answered at all or which survey was theirs because the surveys are returned to a locked box inside the pharmacy.

G. Sample size for surveys: The value of information gleaned from surveys is diminished by low response rates, bias in who responds, too few people being surveyed to be able to tell the difference between sub-groups, and so much variability in response to questions that little can be made of the answers. All of these factors have to be taken into account when deciding how many people to survey. While there are statistical ways to handle all of them, the safest approach is to contact the entire population of interest whenever feasible. When this is done, you don't have to count on inferential statistics to guess how the non-surveyed portion of the population may have actually responded and can concentrate on figuring out how the non-respondents may have felt. If you are doing a follow-up of all 360 patients you treated for disease X over the last three years, it is obvious that you would survey all of them. But, what if you are trying to compare the effectiveness of your treatment (and therapeutic milieu) with that of other therapies performed at your institution on very similar patients who happened not to enter your program. For example, 5,286 patients with uncomplicated chronic classic migraines might have been treated at your institution over the past few years. How many should you survey to find out if your treatment's effectiveness is different from the other approaches? In spite of the expense of conducting and analyzing over 5,000 surveys, I would survey the entire population if at all economically possible because you have no real idea of what other sub-populations there are who may have received very different interventions with very different degrees of effectiveness. If it is impractical to survey everybody, you need to fall back on some rational approach to determining how many people to survey which will probably require application of inferential statistical techniques to guide your decision. These are discussed in the power analysis chapter in the statistics section of the book.

H. Analysis of survey data:

Numerous studies have shown that most survey responses are from those displeased with the outcome or who perceive that they have something to gain from participating in the survey. The second largest group of respondents tend to be those who liked the

outcome. Thus, the data must be analyzed using a model which predicts a skewed distribution of responses. I always use a “worst case” scenario in my analyses. I assume that those not answering would have answered either one way or the other. For example when asking a population of amputees if they have phantom pain, if there is a fifty percent response rate and half of the respondents report phantom pain, I state that at least 25% of the population has phantom pain but that the rate could be as high as fifty percent.

Just as when conducting any study, you presumably had some questions you wanted to answer when you designed the survey and only asked questions which related to those answers. This infers that you have one or more hypotheses that you plan to test and are not just using the shot-gun approach to trying to see if any answer related to any other answer. Your hypotheses will guide you toward which graphic and inferential statistics to use and which variables to test. Regardless of whether you actually planned your analysis in advance or not, patience is the key to an adequate analysis of survey data. The crucial first step is to get an excellent graphic look at the raw data for each question. Averaging the answers can be very misleading because there may be vital clumps of answers along any continuum which will help you understand differences among your respondents.

Survey data are difficult to analyze because the answers to questions do not tend to be independent and many patterns of answers can be present but impossible to spot by eye when questions are analyzed or graphed individually. Inferential statistical techniques can help spot these patterns. Several of the commonly used techniques include:

1. Discriminate analysis and probit/logistic regression - used when trying to see which combination of continuous and discontinuous variables best predicts a categorical outcome measure. Discriminate analysis is used when the grouping variable has more than two choices (a, b, and c) while probit analysis and logistic regression is used when there are only two choices (yes/no). For example, if you asked amputees about presence or absence of phantom pain (a yes/no response) you might want to know if questions about severity of stump pain, frequency of urination, which war was fought in, amount of prosthetic use, and amount of stress the respondent is under predict a yes or no response to having phantom pain. The analysis might tell you that very frequent stump pain and high prosthetic use tend to be related to report of phantom pain while low stress levels and having been in WWII are predictive of not reporting phantom pain. Frequency of urination might not be noted as helpful in predicting whether phantom pain was reported or not. Discriminate analysis would be used if you asked whether the phantom pain was burning, shocking, or cramping (3 categories) in order to see if any of your variables could predict which category a particular patient fell into.

2. Stepwise regression logistic regression are similar to the above tests but are used with constantly varying measures to look for patterns of answers across the entire questionnaire which can differentiate between sub-groups of interest. Continuing the above example, you might use stepwise regression to see which variables predict high vs. low levels of phantom pain when the respondents rated their pain on a scale of zero to ten.

3. Canonical analysis is used with discriminant analysis to reduce the number of variables which predict the classification by determining which variables change together (thus, you only

have to look at the one which gives the very best prediction). This technique usually has a graph associated with it which shows how well the variables differentiate between the two outcome possibilities and highlight outliers which might be going the other way.

4. Principal components analysis would be used to figure out which variables change together and, thus probably measure the same concept. Cluster analysis looks for patterns of responses which could lead to identification of distinct sub-groups of subjects.

These techniques are discussed and exemplified in the statistics section of this book.

I. Critical construction and evaluation of psychological test instruments

A “test” is a “standardized situation in which a person’s behavior is sampled, observed, and described”. This can include recordings of a person’s physiological and emotional reactions. Where do the questions for tests come from? (Ways tests are derived / constructed): In the “rational approach” you ask questions you want an answer to – the question is obviously related to the topic – e.g. how depressed are you? on a depression scale. In the “empirical approach” you use a question because it helps differentiate between groups but no rationale for the question – e.g. MMPI. Questions can also be derived from performing factor or item analysis on some set of responses.

Types of interpretations:

Raw scores (how many math questions correct)

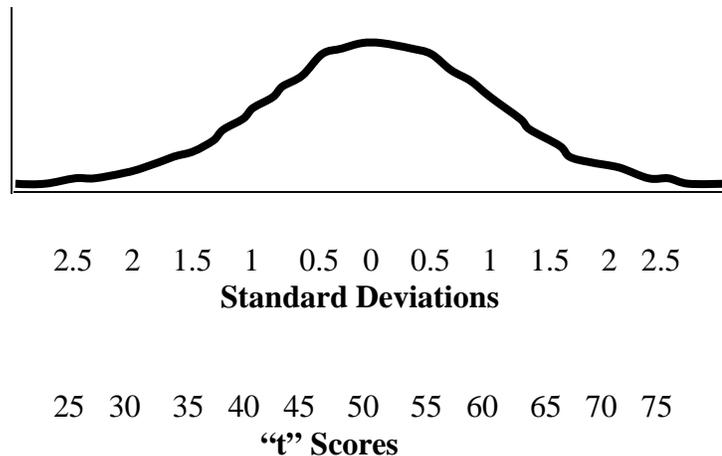
Derived scores relative to norms (e.g. mental age) NOTE: The norms must be relevant to your subject!!!!

Configuration or pattern of responses relative to norms

Clinical or actuarial interpretation

Standardized scores such as “t” scores: Many tests provide the subject’s relationship to the normal population bell curve for each test scale by giving a “t” score. This is just a normalized reference to standard deviations above and below the mean. For instance, a “t” score of 50 means that the subject scored the same as the average of the reference group (normals, a particular kind of nut, etc.) who took the test. The relationship between the bell curve, “t” scores, and other standardized scores is shown in the next figure.

“t” scores and the population bell curve:



Factors effecting reliability of the test:

Presenter – subject interactions.

The longer the test, the higher the reliability up to a point.

Ability of the subject to understand the items on the test.

Characteristics of a good test:

Standardized

Good norms / reference groups

Quantifiable

Reliable

Valid

Efficient in terms of time involved, portability, and expense

Important points and limitations

The test must have been normed on a population very similar to your patient or it's results can not be interpreted.

Tests do not measure without error -- "True score" = "Observed score" + "error"

-- sources of error include time. mood. test content, scoring, departures from standardized procedure, etc.

Only one source of information about a person \not free from bias and distortion.

Involve interpretation which can represent a speculative leap.

Samples of behavior are assumed to be generalizable and to predict or at least be related to real-world behavior.

Interaction between person who is being evaluated (pre-existing characteristics/traits); current life situation (specific symptoms/illness, life stressors, compensation/litigation. etc.); and the specific context of the evaluation situation.

Some of the significant issues in test construction:

Repeat reliability

External validity (does the test actually measure / predict what it is supposed to)

 predictive validity (ex: SATs)

 real world validity (ex: IQ)

Internal validity

Norms

Clarity of questions

Fairness of questions

Evaluating a test's ability to give you the information you need:

An ideal test would only give a positive result when the person was actually sick (a true positive) and a negative result when the person was actually healthy (a true negative). However, tests aren't perfect so they miss in both directions at least to some extent by giving false negatives (the person is actually sick but the test fails to show it) and false positives (the person is actually healthy but the test declares for sickness). Before using any test, you need to know how accurate it is. Accuracy is determined by the test's predictive value and efficiency. We will go into detail on the concepts later in the course.

Efficiency of a test's ability to correctly classify or predict

Diagnosis by test	Actual Diagnosis	
	Real	Real
Positive	True	False
Negative	False	True

Sample Study

Look at the power analysis in the full study.

The raw data results of the pilot are presented in the full protocol
Perform your own power analysis on the data.

Do you agree with what the investigators did?

Chapter 21

Defensive reading of clinical literature - can you trust the subjects and data?

A. What to look for in the "subjects" section of an article:

1. What is the sample? Is it described clearly?
2. How was the sample selected? Was the method of selection appropriate given the purpose of the study? Was some type of sample bias/selection, or loss perhaps influencing the findings?
3. Is the sample representative of the groups to which the study findings should be applied? If not, how does it differ? What are the consequences of the difference?

B. The raw data:

1. How much raw data is shown?
2. Did the investigators report any problems gathering the data?
3. Is there any missing data? Given the complexity of the study and the time it probably took to gather the data, is the amount of missing data about right for the amount you would expect to be missed?
4. Are the data very clean with few outliers? Does the amount of variability go along with your experience and other literature in the area?

Chapter 22

Pitfalls in data gathering methodology

A. Common errors in data gathering strategies

1. Pay insufficient attention to establishing and maintaining rapport with subjects. This often leads to refusal to cooperate or to a negative attitude that can reduce the study's validity.

2. Weaken research design by making changes for administrative convenience (this is a very common problem!!).

3. Fail to explain the purpose and tests to colleagues who will impact on subject availability and attitude.

4. Fail to evaluate available measures thoroughly before selecting one. This often leads to the use of invalid or inappropriate measures or ones with such low validity that the study is wasted.

B. Common errors in questionnaire studies:

1. Questionnaire not the best way to gather the required data.

2. Fail to perform a pilot in which a few subjects similar to the patient population to be tested are given the questionnaire and are then interviewed about what they thought the questions meant, what their answers meant, and what could have been asked but wasn't.

3. Use a pre-made questionnaire not appropriate for the population to be tested or for the data required and fail to pre-test it.

4. Questionnaire too long and/or difficult to answer due to poor structure, difficult vocabulary, etc.

5. Ambiguous questions.

6. Fail to check a sample of nonrespondents for possible bias.

7. Use personality inventories and other self-reporting devices in situations in which the

subject might be expected to fake responses in order to create a desired impression to enhance clinical care.

8. Assume that standard tests measure what they claim to measure without making a thorough evaluation of validity data available (content validity).

9. Attempt to use measures you are not sufficiently trained to administer, analyze, or interpret.

C. Common errors in interview studies:

1. The interview is not sufficiently structured to insure gathering required data or too structured to permit subjects to provide good data.

2. All interviewers do not practice enough to insure that the same data will be gathered and items will be rated identically.

3. Fail to establish safeguards against interviewer bias and subjects telling the interviewer what they think the interviewer wants to hear.

4. Fail to make provisions for calculating the reliability of the interview data.

5. Use vocabulary not understood by the subject.

6. Ask for information that the subject would not be expected to have (e.g. details of drug dosages from years ago or information requiring an understanding of medicine).

7. Ask questions the respondent is not likely to answer honestly (e.g. drug use, homosexuality, etc.)

D. Examples of typical problems with data gathering: (Based on ideas in Principles and Practice of Research: Strategies for Surgical Investigators by H. Troidl et al; Springer-Verlag, New York, 1991.)

1. Selection bias: *An orthopedic surgeon has been performing arthoscopic anterior cruciate ligament repairs on soldiers for five years and decides to get an idea of how effective the procedure has been. The physician uses the Army's world wide locator to find as many of the soldiers as possible, and has their current commanders send information on the soldier's physical abilities relative to the knee repair. About half the soldiers located appear to be doing well and about half have some degree of limitation related to the knee problem.*

The major problems with this design are that soldiers with severe problems would have been eliminated from the Army within a few years (at most) of not being able to meet the physical fitness standards and those in occupations requiring considerable mobility would have

been eliminated at disproportionately higher rates than those in relatively physically unchallenging occupations. Thus, the physician would have only been able to locate those with either successful outcomes or with relatively minor impediments who were in relatively undemanding jobs. In fact, the results of the intervention were probably quite poor.

2. Sampling errors: *"A random sample of people registering at hotels was asked about their preference for using general taxes to pay for medical care for the indigent. Great care was taken to insure that all races and sexes were representative of the population and that survey sights were distributed in both population and geographic representative patterns. The results showed that the vast majority did not want to support medical care for this population from general tax revenues."* □

Less poor people than wealthy people register at hotels. Thus, there was probably an income gradient among the respondents which is very different from that of the general population.

E. Common mistakes with data gathering techniques: The most common reason studies fail is that data are not gathered accurately and consistently.

1. Objective recording of events: Remember that objectivity is not a human characteristic. **YOU GET WHAT YOU EXPECT.** When methods of taking data are designed, the major effort must go toward distinguishing interpretive from objective data. For example, when observing an interaction between two patients, you could report either "the patients were sleeping" or the patients "appeared to be sleeping and remained in a sleep position for an average of 6.3 hours" The second way permits you to record actions rather than interpretation of actions. This will permit further evaluation of the patients' activities.

2. Problems with data gathering methodology & interpretation: These are the same as the assumptions about prior work discussed earlier on page 122.

Practical exercises for section C

1. Choose a topic for a mail response survey. Provide a short explanation of what you are attempting to learn through your survey. Design the survey with at least ten questions. Some of the questions need to be scales, some yes/no, and some in other formats. Explain how you will choose, test, and finalize the questions. Provide the survey in user-friendly format.

2. Choose a topic for a longitudinal study. It can be experimental or observational. It can also be entirely prospective or part prospective and part retrospective. Provide a short explanation of what you are attempting to learn through your study. Discuss the strengths and weaknesses of the design vs. cross-sectional studies. Include a description of the typical types and uses of longitudinal studies. Provide a clear description of the study design to include the specific outcome measures you will use to track changes over time.

3. Choose a topic for a cross-sectional study. It can be experimental or observational in nature. Provide a short explanation of what you are attempting to learn through your study. Discuss the strengths and weaknesses of the design. Include a description of the typical types and uses of cross-sectional studies. Provide a clear description of the study design to include the specific outcome measures you will use to assess changes over time.

4. Either make-up a study or use one from questions 2 or 3 above which involves collecting highly subjective data. Briefly describe the study, the data, and what you are trying to learn. Specify how you will harden the data – insure that you cover, but do not limit your answer to, the techniques for hardening subjective data discussed in the text.

Section D

Statistics for evaluating the clinical literature & interpreting clinical data

Chapter 23

Concepts of clinical data analysis

The motto for this section is GARBAGE IN = GARBAGE OUT

A. Distortions: This section is about how to interpret data. Everyone interprets data continuously. We get through life by assigning meanings to our perceptions (data), associating them with our current experiences, and making a judgement on the perceptions based on that association. The problem with this method when applied to clinical data is that we make very heavy use of our background knowledge to interpret what we are seeing. This frequently leads to distortions in our conclusions for several reasons including:

1. Faulty judgement of treatment success rates because disproportionately more failures do not return and failures are more easily forgotten than the shining successes.
2. Attribution of symptom attenuation to the treatment rather than placebo effects, random variation, natural decreases with time, etc.
3. Selectively interpreting the data based on expectations and / or outcomes.
4. Inaccurately associating symptoms with outcomes. Many people have many symptoms which co-vary in complex (or independent) ways. It is all too easy to see a few apparently similar patients with similar problems and mistake symptoms unrelated to the problems which happen to appear in most of the patients as being related to the disorder. This is commonly done when headache is added in to the symptoms of hand pain syndromes.

B. Appropriate use of statistics: Statistics is a very complex field. You can support any meaningless relationship with some test. Statistics in clinical areas are only used (1) to help you determine whether trends that you see are likely to be consistent and repeatable and (2) to find patterns in complex data (such as surveys with many questions) which are not easily seen by eye. “Statistical decisions” is a nonsense term. You use statistics to guide you toward a decision.

C. Analysis is not equivalent to inferential statistical analysis: The worst mistake you can make in analyzing your data is to stuff it into a computer's statistical analysis package before looking at the raw data. Very careful, slow, deliberate examination of the raw data from scatter plots which show the distribution of the raw data (not graphs with means) will show you the relationships between the variables and, very importantly, the shapes of the distributions of the data. This is your only way to look for clusters of data and relationships between apparent outliers and the rest of the data points. The human eye is the most sophisticated pattern detection system available. Due to the extreme variability of human data, the statistical packages frequently miss important relationships you will see instantly. Graphics packages capable of plotting individual raw data points from your data entry package are commonly available. You will probably get much more from a careful evaluation of your raw data than you will from any amount of statistical printouts.

D. When to choose your analysis: You should determine approximately what kinds of analyses you are going to attempt while planning the design of the study (1) to insure that you can actually analyze the data you intend to collect and (2) so you collect it in a format optimally compatible with analytic methods. Many studies gather data they can never use.

E. Selection of statistical tests: There are so many mistakes in the literature that you can not trust the design strategies used by published studies similar to your study. If you have not had (and remember and understand) a course in basic statistics, you should read a basic book on **clinical** statistics (the others may be quite misleading) so that you can communicate with a biostatistician who is specifically experienced in **clinical experimental design and analysis**. Biostatisticians without clinical training do not understand patient individuality and are more likely to recommend the wrong tests. If you decide that you are sophisticated enough in statistics

to choose a test, read up on it the way you would on a medical procedure prior to using it. There may be many very helpful tests you have never heard of.

The microwave oven simile:
An explanation of
the orientation toward statistical understanding presented in this book

Let's say you need to use a microwave oven. To get started you need to be able to get it to do what you want it to do. It might be interesting to have a one second overview of how a microwave works but you really don't need to know what it is doing, or how it is doing it, to use it. You certainly don't need to know the theoretical physics and mathematics underlying it. You don't need to know the structural engineering that went into it. But, you'd better know enough about what it is doing to convince you that you could be seriously harmed if the door interlock fails.

Microwave ovens can ruin your food if you use them incorrectly. If you understand their limitations and use them correctly, you have a fair chance of having your food come out optimally. Because microwaved food doesn't always look cooked the way we are used to, sometimes you can't tell if it is done correctly until it's too late - and you find that you have prepared yourself a real mess to eat - and a worse one to clean-up.

Microwaves can be pretty complex to use. At first they are confusing and you can't predict what they will do. You are *apprehensive* about using it and would rather not go further into its instructions than you have to in order to cook whatever you want to. However, the more you know about your microwave oven's capabilities, the more you can do with it. To learn more, you have to read the instruction manual and try it out. The more you learn and the more you try, the better you get at using it and the more comfortable you become. After awhile, it becomes another familiar tool which you can confidently use to accomplish whatever you can use it for.

To cook food in a microwave, you need to have one available. For people who need to analyze data, that means getting a statistical software package of your own or getting access to one.

Microwave ovens differ from each other in complexity, ease of use, and capabilities. If all you want to do with a microwave is heat water for your tea, you can get a small, easy to use, inexpensive oven. The corresponding statistical package would be something like "Mystat." It will do very simple power analyses, simple statistical tests for simple designs, and help you characterize your data. But it doesn't have a built in spreadsheet and it can't handle complex situations or warn you when there are crucial holes in your data.

If you want to defrost a roast and then have it cooked to a certain temperature while being browned, you need a pretty complex microwave oven with a convection oven built in. It will be more difficult to learn to use, have more complex directions, and probably cost quite a bit more for the extended capabilities. So, if you want a really good database program built into a statistics package that can do anything you are likely to ever need, expect to spend more time, effort, and money on it. Examples of these would be "Statview" and "SPSS-PC."

You have to choose a microwave oven you can get along with. Some microwaves with few capabilities are far more complicated to use, have masses of bewildering buttons, and are more expensive than some with far more capabilities which are relatively simple to use and don't have lots of unnecessary buttons. Some people like one presentation while others like a different one. Before you run out and buy a statistical program, try a few and see which one goes along with your way of thinking. For example, Statview and SPSS-PC approach moving the data around very differently. Some people never do catch on to how they need to move the data in Statview, while SPSS-PC is more direct but more cumbersome to use.

One word of caution: You can't use the microwave to freeze the carrots. You need to recognize the limitations of your tool and use it appropriately. If you want to crush garlic using the microwave oven, you would have to pick it up and use it as a hammer. It would do a lousy job and probably break. Statistics packages can't transform bad data into good data and they usually can't tell you when you are using the wrong test. But, they can put out endless masses of meaningless analyses which sure look impressive.

Chapter 24

Descriptive statistics for evaluating clinical data

A. Overview of the data: Once your data have been gathered, the worst thing you can do is cram them into a statistics package. Instead, you need to get an overview of your data's distribution so you know which tests (a) are likely to be useful (not to mention valid) in helping you understand your data and (b) to use to test hypotheses you generate when looking at the raw data.

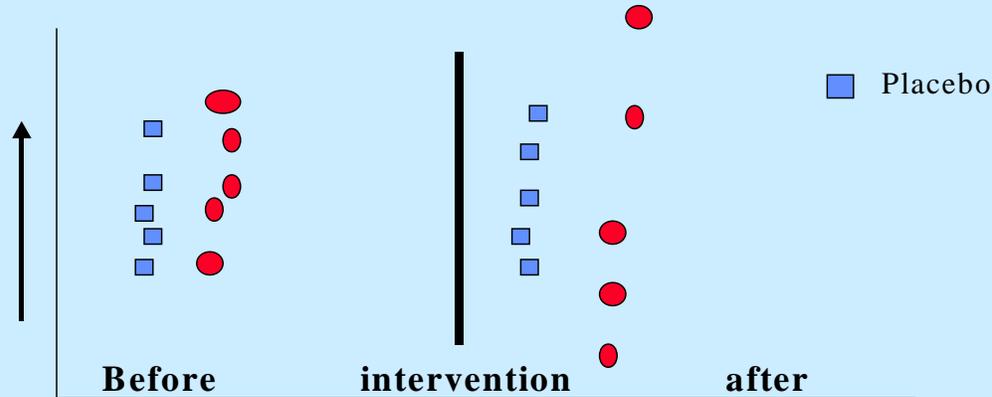
B. Variables and variability:

1. The importance of recognizing variability: You must have an excellent grasp of how variable your responses are or you will not know what happened. Very often, the only result of an intervention is increased variability in levels of the outcome parameter because some subjects react to the intervention while others do not. It is also quite common that there is so much variability that a perceived response to the intervention or an apparent change over time is simply a random movement of numbers within the limits of normal variation. The pattern of variability is crucial because sub-groups may be obvious when variability is evaluated which would be obscured by lumping the data together into means.

The other reason you need to look at variability is that weak interventions frequently only produce changes in variability. This is because some subjects don't respond at all while a very few who are hypersensitive to the intervention respond a lot. Thus, the average may remain virtually the same but the variability increases tremendously. This change is illustrated in Figure 4.

Figure 4

Very often, the **only** result of an intervention is **increased variability** in levels of the outcome parameter because some subjects react to the intervention while others do not.



2. Types of variables:

- a. Dichotomous variables: Only two choices are available. E.g., yes/no or true/false.
- b. Nominal variables: You have several choices (A, B, C) presented in random order as there is no grading (highest to lowest) associated with the choice of which variable is A, which is B, and which is C. For example, which do you like best: Coke, Pepsi, or Cod Liver Oil?
- c. Ordinal / non-parametric variables: These variables are in a scaled order usually from low to high. Examples include "Likert" scales and ratings from little, through medium, to a lot. For example, "rate your pain on a scale of zero through ten where zero is no pain and ten is so much pain you would faint if you had to bear it for one more second," or "rate how much you like cooked carrots on a scale of 0 - 5". There is no way to tell if a rating of two is really half that of four. In other words, you can't count on equal distances between the ratings.
- d. Continuous / parametric variables: These variables include such measures as weight and height. They have true zeros and equal space between points.

C. Don't forget the basics: When you are ready to begin reducing your data, make sure you do not make elementary mathematical mistakes which won't be noticed during the graphic and complex analyses to come. For example, percentages used as outcome variables foul up too

many people:

Percent change = change / initial score X 100. If the initial score is 8 and the current score is 6, the percent change is $-2/8 = -25\%$.

D. Importance of "eye-balling" graphs of the raw data:

1. Scatter plots: The human eye is better at picking patterns out of GRAPHIC data than any statistical technique. When you look at your data displayed in graphic format, you will quickly spot trends and outliers that the computer will never tell you about. You will also see problems with distribution which would make such measures as "means" misleading.

Figure 5 shows a typical scatter plot in which every data point is shown. Note that at least one of the numbers is very far away from the others.

Figure 5

Scatter plot showing raw data



2. Frequency distributions: Imagine the raw drug response data from a three dose group study looking like the illustration in Figure 6. Application of standard statistical techniques would result in your missing the change in distribution as a few outliers reacting at low doses grow in frequency at medium doses and become the majority at high doses. This is the way many studies turn out that are incorrectly reported in the literature. The authors never looked at the raw data and have no idea what happened.

Figure 6

Dose - Response curves for three doses of a test drug



Figure 7 illustrates common distributions of data. The left column depicts variables that change continuously, such as weight while the right column shows variables that are discontinuous such as eye colors (without gradations). On each graph, the horizontal line (abscissa or “x” axis) shows the scores or categories while the vertical line (ordinate or “y” axis) represents the number of subjects producing the number. The top set, called normal (or bell shaped), is rare in clinical work. The lower two are the most common. Unfortunately, most common statistical tests and measures of dispersion require normally distributed, continuous data for their results to be meaningful.

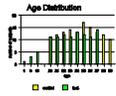
Figure 8 shows the age distributions for the control and test groups in a drug trial. It is obvious from looking at the graph that the distributions are very different. However, the means would be very similar and the few low numbers for the test group would not show particularly in the standard deviation. Thus, a typical statistical test is very likely to miss the difference.

Figure 7 **Common distributions of data**
 (Number of subjects responding at each point is shown on the vertical axis)

	Discrete	Continuous
Normal Distribution Bell curve		
Bimodal		
Skewed		

Figure 8

Example of skewed distribution of data



The axes of frequency distribution graphs do not have to be evenly dispersed arithmetic progressions. For example, the horizontal axis could still represent continuous data such as blood pressures but be divided into intervals to clump the data into meaningful segments so it is easier to evaluate by eye. The vertical axis may show the values increasing logarithmically or geometrically rather than arithmetically. This can have a profound effect on how the data look and how one might interpret their meaning. Figure 9 illustrates a set of data shown in intervals with a logarithmic frequency axis on the left and continuously with an arithmetic frequency axis on the right.

Both axes may be shown with breaks in them if there are long distances between points - watch out for the practice of showing such breaks and for not having the origin (the lower left corner where the two lines intersect) may not be 0,0 which could lead to misunderstanding the amount of change represented.



Figure 9 **Examples of different axes**

Log - clumped

Arithmetic - continuous

E. Indications of central tendency:

1. Mode - most frequent - not sensitive to outliers
2. Median - The exact middle score or value in a distribution of scores; the value (the 50th

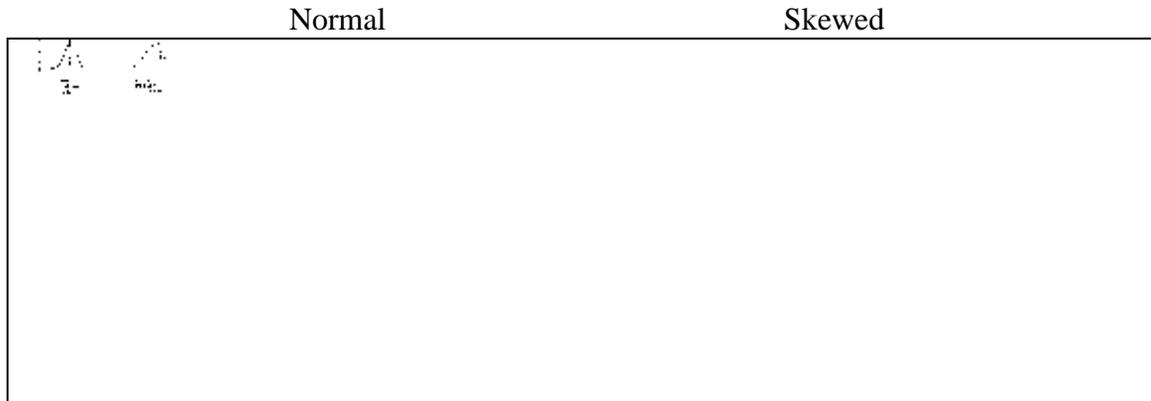
percentile) above and below which 50% of the scores lie. Not very sensitive to outliers.

3. Mean - average - balance point - super sensitive to all scores, especially outliers. The mean is usually indicated by an "x" with a bar over it.

Crucial note - the "mean" of a group of numbers is meaningless if even one of the numbers is a far outlier. For example, the mean of 1, 3, 5, 7, 9, 11, and 13 is 7 but the mean of the same numbers with 80 substituted for 13 is 17.

Figure 10 illustrates the effect of a skewed distribution on the mean, median, and mode.

Figure 10 Effect of a skewed distribution on the mean, median, and mode.



F. Measures of dispersion and variability

1. Range: distance (number of points) between the lowest and highest scores. Range is frequently given as the low score and high score rather than just the distance. So, if the low score is two and the high score is six, the range is four but people frequently write "range = 2 - 6" so the reader will have the additional information.

2. Variance: Gives an idea of how much dispersion of individual points there is around the mean for the group. The distance of each score from the group mean is squared and then all

of the squares are added. For example:

<i>score</i>	<i>group mean</i>	<i>individual variance</i> (<i>mean - score</i>)	<i>variance²</i>
2	6	4	16
6	6	0	0
10	6	-4	16

$$\text{Variance} = \frac{16 + 0 + 16}{3} = 10.7$$

3. Standard deviation: square root of variance. Please note that the "Standard Deviation" (SD) is not magic. It is only a measure of variation and quickly becomes meaningless if it is not applied to a normally distributed group of numbers. Standard deviations are very sensitive to the exact position of each score and, thus, to outliers. SDs are usually shown after a mean. For example: 27 ± 7.2 or $27 (7.2)$.

In a normally distributed population:

- The group mean ± 1 SD contains about 68 % of the scores.*
- The group mean ± 2 SDs contains about 95 % of the scores.*
- The group mean ± 3 SDs contains about 99.7 % of scores.*

THUS, when the results of a test show one score (such as a blood value) to be one SD away from the mean, remember that 32% of the **normal** scores are expected to be that different from the mean. The value is not automatically "abnormal."

4. Coefficient of variation: This underused but very valuable measure is excellent for comparing variations when the means are very different because the coefficient of variation is the standard deviation divided by the mean multiplied by 100. Even if the amount of variation in two samples is proportionately the same, the standard deviations will be very different if the means are very different. For example:

sample one: 2, 4, 6 mean = 4, SD = 2
 sample two: 10, 20, 30 mean = 20, SD = 10

In reality, the variabilities in the above example are the same but the standard deviations make them look very different. This is demonstrated by calculating their CVs as follows:

$$\text{for sample one: } CV = \frac{2 \times 100}{4} = 50$$

$$10 \times 100$$

for sample two: $CV = \frac{\text{-----}}{20} = 50$

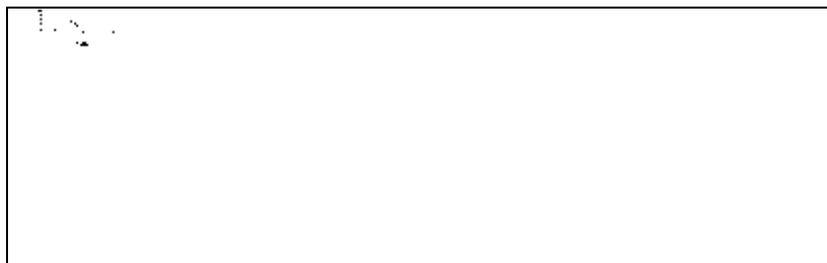
Thus, the two groups actually have the same amount of variability. This becomes important because (a) many statistical tests, such as “t” tests can not be used unless the groups have similar variability and (b) changes in variability are an important indication that an intervention has had an effect.

5. Percentiles: Percentiles are a measure of relative standing in a group. Raw numbers generated by any kind of test tend to be distributed quite randomly in the real world and it is frequently difficult to tell where any one of them sits in relation to the distribution of the others. If you have a test that has produced many numbers ranging from 10 to 50 and your latest run of the test produces a 43, how do you know what proportion of the scores are above and below it? You might also want to divide your scores up into the top ten percent, bottom 10 percent, etc. so you can get a feel for the relationship of all your scores to each other. The percentile rank of a number is the percent of scores less than the score you are interested in. You could think of the median of the scores being the score that corresponds to a percentile 50%. The remainder of the scores can be divided up into quarters to make evaluating their relative standing easier. The upper quartile point is the score that separates the top quarter of the scores from the lower 3/4 while the lower quartile point is the score that separates the bottom quarter from the upper 3/4. The interquartile range is the point spread between the upper and lower quartiles - which gives you a good idea of the amount of spread around the median.

To figure out how many scores would be in the top or bottom X percent of the numbers recorded, multiply “X” by the number of scores. For example, if you recorded 60 numbers and want to know how many are in the top 10 percent, multiply 0.10 by 60, which = 6. Thus, the top six numbers are in the top ten percent.

6. Weakness of descriptive statistics: Using descriptive statistics without looking at the raw data can cause the investigator to entirely miss the idea of what the results show. For example, Figure 11 illustrates the mean of a bimodal distribution.

Figure 11 Location of the mean in a balanced bimodal distribution



Chapter 25

Probability and significance testing

A. Probability: Probability is a fancy way for saying “the odds”.

B. Chance Events are independent: Keep in mind that if all results have an equal chance of occurring, the chance of any one happening in one trial is $1/\text{total \# possibilities}$. The chances of getting two "1"s from the same die thrown twice or two dice thrown once are $1/6 \times 1/6 = 1/36$. This holds for betting a "1" on the first trial and a "2" (or any other number) on the second trial. Please also remember that random events occur randomly so they can form clumps such as ten heads in a row coming up when tossing a coin into the air many times. This is the same explanation for “runs” in dice when an unlikely combination turns up many times in a row.

C. Levels of probability: The term “0.50 level of probability” means the result is likely to happen one way 50% of the time and the other way 50% of the time. A 0.05 level of probability means the result is likely to happen one way 5 times out of 100 times the test is repeated and the other way 95 times out of 100 times the test is repeated. So, if you repeat your study 100 times, it should come out the way it did when you got a 0.05 probability 95 of the 100 times. **REMEMBER** that it *probably will come out significantly differently 5 of the 100 times and that the next time you do the study could be one of those five times!!*

D. One vs. two tailed probability: This is the concept that it is twice as hard to guess which way your groups will differ from each other than it is to guess that they will be different in one specific direction (e.g. group one will get better faster than group two). Curves which estimate the probability of an event look like bell curves with very long tails at either end. It is much more likely for an event to be in the middle (where it is tall and fat) than under one of the very low tails sticking out at the sides. If the probability that an event will be high or low on the curve is equal, then it has an equal chance to show up under either tail. To put this into "English": If all you care about is that the groups are different from each other (you do not care which group gets better first as long as one or the other does), then there is theoretically as much chance of either group getting better first so you could get significance either way - from either "tail" of a curve showing likelihood of occurrence rates (a probability curve). If your hypothesis is that group one will get better faster than group two, you lose your bet if it turns out that group two gets better significantly faster. Your predicted result is only on one tail of the probability curve. So, you have half the odds of group one getting better faster than of the groups simply being different. In probability terms, an answer is two tailed because you do not care about direction. The same result from a statistical test giving 0.05 probability with a two tailed question is 0.025 for a one tailed question. Thus, if your cut off for declaring a difference to be significant is 0.03, you

would win if it was a one tailed test but loose if it was two tailed. Many investigators do not understand the difference and just pick whichever shows that their work was "significant."

E. Decisions on level: *You have to decide if a level is good enough for your needs!* Can you accept the chance that you will be wrong five percent of the time? Can you permit five percent of your outliers to die unpredictably if they are humans? If there is a 95% chance a test will be positive if you have cancer, can you afford the 5% chance of missing it?

F. Multiple testing of the same data: If your "significance level" is 0.05, then 5 times out of 100 (or 1 out of 20 times) you should expect to find "significance" by chance alone. Thus, if you have a study with 5 groups and you test all of the combinations of the five groups (1 vs. 2, 1 vs. 3, 1 vs. 4, 1 vs. 5, 2 vs. 3, 2 vs. 4, 2 vs. 5, 3 vs. 4, 3 vs. 5, and 4 vs. 5) you will have 10 tests. If each group was evaluated on two variables, you will have done 20 tests. The same number of combinations are accrued if you have two groups with five outcome variables. One of those tests should be "significant" by chance alone. Thus, studies with multiple variables usually find "significance" someplace but their results are rarely replicated by other studies because they only occurred by chance. There are special tests which avoid the multiple testing problem.

G. Distribution of random events: Random events tend not to be normally distributed. Instead they tend to (1) clump and (2) have unusual values which return to the mean when the group is retested. Thus, a clump of random values can easily be misinterpreted as a real finding which disappears when the study is replicated. Any time you flip a coin to look for the number of times you get heads or tails, you are very unlikely to get an even number of heads and tails even if you flip a hundred times. You are also very likely to get runs of all heads and all tails. This frequently happens in epidemiological studies when a disproportionate number of people in one tiny area come down with some misery. It can really be a random event due to clumping of individual random events but it is next to impossible to get anyone in that town to believe it. Blaming a government cover-up is much more satisfying explanation.

H. Testing for "significance":

1. What are significance tests?: Significance tests are the infamous "odds" making tests found misused throughout the literature. They are only good for letting you know the odds that your outcome is not random. They can't prove anything or make a decision for you.

2. Statistical vs. clinical significance: A very common comment about an analysis of data is that the results were statistically significant but not clinically significant. It is up to you to choose the correct test - which will look for the changes you are interested in at the level of significance and type and magnitude you want to know about. If you need help finding the test, why not get help? Using a test for very small differences in central tendencies when the clinical significance is indicated by a change in the relative number of extreme outliers does not make sense but is done every day.

I. Assumptions in the use of “significance” tests: Each test is valid only if your data meet the entrance criteria (Assumptions) the test is designed around. Unless you are one of the very few biostatisticians who actually understand the theory behind the tests, to disobey a clearly presented assumption for use of the test, is to invalidate the test's results. This is most commonly found when the "t" test is used for clinical data. It demands that the two groups be very similar in variation and be normally distributed. It is rare indeed that clinical data are normally distributed or that two groups of humans have similar distributions on much of anything.

J. Repeated measures vs. independent groups: If you tested the same subjects two or more times, most of the changes in the numbers should be due to factors such as your intervention and time between the tests. Since you have "repeated measures" from each subject, you can use "paired" data tests. Their formulae are very different than those for independent groups and it is easier to get significant differences because of the assumption that the differences between the readings are not due to random differences between people.

K. Parametric vs. non-parametric statistics: If your data are parametric in nature (continuous, real numbers having set, equal distances between them and with a real zero - e.g., weight or height), you can use tests designed for this kind of data (such as “t” tests) as long as your data meet the other assumptions for the test (most importantly that the two groups have very similar variability and that both distributions are normal). If your data is non-parametric (no real zero, not necessarily equal distances between the numbers - e.g. rating scale answers such as “rate how much pain you are in on a scale of 0 to 10 - where the distance <intensity difference> between 0 & 1 might be very different than between 9 & 10”), you need to use tests for this kind of data. You will get the wrong answer if you use a test set for parametric data.

It is usually better to use parametric tests over non-parametric ones if your data meet their stringent entrance criteria. This is because it is easier to reach levels of significance with parametric tests than non-parametric ones. Non-parametric tests do not have as stringent entrance criteria as parametric tests so need to be more conservative about declaring odds ("less powerful" in statistical jargon). For example, if a "t" test needs 91 samples to find significance, a "U" test will need 100 to find the same level of significance for the same data when all other factors are the same.

L. Degrees of freedom: The number of “degrees of freedom” in the structure of your study refers to the number of sample values that cannot be calculated from knowledge of other values and a calculated statistic. For example, if you know the mean of a group of numbers, (e.g., by knowing a sample mean), all but one value would be free to vary. If there were ten numbers in the group, there would be nine degrees of freedom (DF).

Chapter 26

Decision/Risk Analysis and Evaluating Relative Risk

A. Decision Theory:

Even before you get started you need to decide whether its worth doing the study. Are the odds of success high enough for you to commit the time and resources? Which approach to answering your question is the one most likely to lead to a strong answer?

In general, there are three ways of looking at how certain you are of something: (1) you may already know for sure that it will or won't be, (2) you may have a good idea of what the odds are that something will happen, or (3) you may have no way to predict. If you think of this in terms of probability, certainty that something will happen equals one while certainty that it will not equals zero. You are taking a risk any time you work with an unknown probability. The amount of risk / uncertainty you are willing to accept is somewhere between just below one and just above zero. A series of steps have been developed to help chose the right course when you don't really know what the odds are.

1. The “payoff” matrix: Let’s say that you have an objective you wish to accomplish such as finding out if biofeedback cures chronically frozen shoulders. You would decide on several methods for making this determination (e.g., chart review, open study, controlled study) and determine what the likely outcomes of your investigation might be (few changes in symptoms for anyone, some improvement for a few patients, most patients cured). Next you would construct a matrix such as that shown in Table 9 in which you would assign values to the combination of events you might see happen in relation to their importance to achieving your objective on a scale of one (low / minimum benefit) to ten (high / maximum benefit). When assigning values, you have to trade off the amount of resources you invest against the value of the outcome. Each row of the “payoff” column contains the worst and best scores from the “potential outcomes” column. You would accept the payoff row that minimizes risk - the one with the highest score on the worst possible outcomes (best of the worst) relative to the lowest of the best possible outcomes (worst of the best). In the table below, the open study row represents your best bet for balancing resources with potential findings.

Table 9 **Payoff matrix**

Power of study design	Potential Outcomes			Payoff	
	little effect	some effect	many cures	worst that could happen	best that could happen
controlled	1	6	10	1	10
open	5	7	6	5	7
chart review	8	5	3	3	8

In reality, you would usually have some idea of what the probability is that the treatment will work to a certain extent. Let's say that you feel that there is a 30% chance (probability of 0.30) that you will cure most of the patients. As “certainty equals 100% (1.00) you have 70% of the odds to account for. You would divide them up as best you can between the remaining two likelihoods so might say that there is a ten percent chance (0.10) of not helping and a 60% (0.60) chance of having a moderate effect on most people. This is a reasonable situation as you wouldn't do the study if you didn't think there was a major chance that you would have more than a minimum impact on the disorder. There is also not that much chance that you will have developed a treatment that can really cure just about everybody. The next step is to include the probabilities in your payoff matrix by multiplying your importance values by the probability assigned to each potential outcome as illustrated in Table 10 below.

Table 10 **Payoff matrix with low probability of great success**

Power of study design	Potential Outcomes			Payoff	
	little effect p = 0.10	some effect p = 0.60	many cures p = 0.30	worst that could happen	best that could happen
controlled	0.1 X 1 = 0.1	0.6 X 6 = 3.6	0.3 X 10 = 3.0	0.1	3.0
open	0.1 X 5 = 0.5	0.6 X 7 = 4.2	0.3 X 6 = 1.8	0.5	4.2
chart review	0.1 X 8 = 0.8	0.6 X 5 = 3.0	0.3 X 3 = 0.9	0.8	3.0

Your choice would still be to do the open study but it is very close to being worth doing a chart review only as a first step. Let's say that you believed that the probability that your treatment would cure just about everybody is relatively high (80%). This would make a radical change in

your matrix as illustrated in Table 11 below. The balance now really favors doing a controlled study because the odds of bombing out aren't really that different between the methods while the odds of succeeding are very high if a controlled study is done.

Table 11 Payoff matrix with high probability of great success

Power of study design	Potential Outcomes			Payoff	
	little effect p = 0.10	some effect p = 0.10	many cures p = 0.80	worst that could happen	best that could happen
controlled	0.1 X 1 = 0.10	0.1 X 6 = 0.60	0.8 X 10 = 8.0	0.1	8.0
open	0.1 X 5 = 0.50	0.1 X 7 = 0.70	0.8 X 6 = 4.8	0.5	4.8
chart review	0.1 X 8 = 0.80	0.1 X 5 = 0.50	0.8 X 3 = 2.4	0.5	2.4

2. Ranking importance: Very often several approaches to solving a problem may be about equally efficacious but differ in relative costs to you. For example, you probably have a limited amount of funds, assistant power, and patients available so need to figure out which approach gives the optimal combination. In the example shown in Table 12 below, approach “B” has the lowest rank score so is the one to use. However, this technique assumes that each factor (cost, person-power, and number of patients) are equally important to you. This may not be the

Table 12 Use of rank ordering to determine which approach is most cost effective (Ranks are given in order of importance with lowest being most important.)

Approach	Cost per subject		person-power needed		number patients needed		Sum of Ranks (lowest is best)
	\$	rank	hours/subject	rank	#	rank	
A	\$10	1	20	3	150	3	7
B	\$20	2	15	2	80	1	5 (best)
C	\$50	3	8	1	120	2	6

case. For example, you might be able to get plenty of patients with the disorder in question and be able to get extra personnel support by putting in a few extra hours yourself but have little

flexibility in funds. Thus, you would want to give more weight to funds, somewhat less weight to technical support (but still some as your free time is probably precious to you), and minimal weight to number subjects required for the study. The example in Table 13 below shows the same example with weights added. When weighted for cost, approach “A” is now optimum.

Table 13 Use of *weighted* ranks to determine which approach is most cost effective
 (Ranks are given in order of importance with lowest being most important while weights are given in order of importance with the highest being most important.)

Approach	Cost per subject Weight = 10		person-power needed Weight = 5		number patients needed Weight = 1		Sum of weighted ranks (lowest is best)
	\$	rank X wt	hours /subject	rank X wt	#	rank X wt	
A	\$10	1X10	20	3X5	150	3X1	28 (best)
B	\$20	2X10	15	2X5	80	1X1	31
C	\$50	3X10	8	1X5	120	2X1	37

3. Patient treatment decisions: The weighted ranks approach is frequently used to help patients decide which treatment option is best for them. The example in Table 14 below illustrates that sometimes an expensive, relatively risky surgical procedure is better in the long run than doing nothing or a relatively benign medical procedure not likely to have a substantial impact on the problem. In this example, probabilities could be used instead of weighted ranks.

Table 14 Use of weighted ranks to determine which treatment would be most likely to produce the most desired results

(Ranks are given in order of importance with lowest being most important while weights are given in order of importance with the highest being most important.)

Approach	long term cost of treatment Weight = 1		likelihood of negative outcomes (crippled by surgery, diseases from not being treated, etc) Weight = 10		likelihood of positive outcomes (longer, healthier, more active life, etc.) Weight = 5		Sum of weighted ranks (lowest is best)
	cost in thousands	rank X wt	odds of permanent negative outcome	rank X wt	odds of permanent positive outcome	rank X wt	
no Rx	\$0	1X1	0.7	3X10	0.1	3X5	46
medicine	\$10	2X1	0.1	2X10	0.3	2X5	32
surgery	\$50	3X1	0.2	1X10	0.6	1X5	18 (best)

B. Evaluation of Relative Risk

This is not a course in epidemiology. However, so much of junk science is based on misuse and misunderstanding of the basic principles of epidemiological evaluation and so many clinical studies fall apart when they touch occurrence rates that I am covering just a touch of the principles here.

Epidemiological studies usually calculate the “relative risk” of getting sick. This is a statistical calculation of differences in disease rates of different groups of people. Relative risk calculated as the difference between the risks that each group has of getting the disease. People from each group are divided as follows between those with and without the disease:

	Disease	No disease
Group 1	A	B
Group 2	C	D

Relative risk = (A: # in group 1 with disease divided by C: # in group 2 with disease) divided by (B: # in group 1 without disease divided by D: # in group 2 without disease).

The answer says how many times more at risk group one is than group two.

A relative risk of 1.0 = No difference in rate of disease between two groups (both groups have the same chance of getting sick).

A relative risk of 2.0 = Group 1 has twice the risk of group 2 (group 1 is 100% more likely to get the disease).

A relative risk of 0.5 = Group 1 has half the risk of group 2 (group 1 is 50% less likely to get sick (etc.) than group 2).

An example should help clarify this concept. Let's say we want to know how menopause affects the likelihood of women getting migraines, we would do a survey of women which includes questions on whether they have reached menopause (eliminating people with hysterectomies, etc.), and whether they have migraines (we would include the characteristics of the headache rather than the word "migraine"). Note that the survey has to be large enough for us to have confidence in the rates. The following is a fictitious set of numbers from this imaginary survey:

	Migraines	No Migraines
Pre-menopause	A 1,500	B 3,000
Post-menopause	C 150	D 10,000

We would calculate the relative risk as follows:

$$(1,500/150)/(3,000/10,000) = 10/0.3 = 33$$

So, women have about 33 times the risk of having migraines before menopause than afterwards (remember that these are imaginary #s).

Many epidemiologists feel that, due to the normal relatively low reliability of rates from small surveys and instability of human data, increases in relative risk from 1.0 - 2.0 or decreases to 0.5 should be viewed with great caution. Some statisticians feel that any number under 3 is not to be trusted unless the survey was absolutely huge (millions of people) and that the rates are very stable. Thus, you need to be very cautious when interpreting relative risk statistics.

Chapter 27

Power analysis - determining the optimal number of subjects

A. The need to establish an estimate of required sample / group size: When designing a study it is vital to be able to estimate how many subjects you will need. Accurate estimation of the optimal number of subjects required to correctly perform a proposed study is critical to decisions regarding approval of that study. It is frequently the case that more subjects are required to reliably determine the presence or absence of a clinically important difference between two groups or between the study group and the known population than are available for participation in the study during the time the investigator will be at the institution or are logistically and economically feasible to run through the protocol. Underestimation of the number of subjects required leads to gathering insufficient data and having to redo all or part of the study long after the study was thought to have been completed. In those cases where batch analysis of samples is required for consistency, it may be necessary to repeat the entire study. Overestimation of the number of subjects required can lead to the decision that the study is too costly in manpower and time to perform at all. If it is attempted, and more subjects participate than are required to achieve a consistent, reliable difference, considerable time on the part of the investigators and of the unneeded subjects is wasted and the funds spent performing the extra tests are lost. Army regulations require that you not attempt a study for which not enough subjects are available to have an excellent chance of finding a difference between conditions because you would be wasting your and your subjects time. Similar Army regulations prohibit you from running more subjects than necessary to prove your point to avoid wasting time and resources on the extra subjects.

B. How to estimate required numbers of subjects:

1. The minimum number of subjects likely to be required to differentiate between two groups or between a study group and the general population can be determined statistically through well accepted power analysis techniques if the variability in the data being collected is known for the group(s) under study and/or the population with which the group is being compared are known. In general, the greater the variability, the more subjects will be required to determine a meaningful clinical difference.

2. Balancing interpretive errors: A “Type I” error is to say that a relationship exists when there is none while a “Type II” error is to say that there is no relationship when there is. If you do not have sufficient subjects in your trial, you are quite likely to miss an existing relationship (a Type II) error because the relationship will be obscured by noise from intersubject variability.

You may also declare a relationship to exist when there isn't one (a Type I error) because a tiny sample happens to randomly sort itself out so high numbers within the limits of normal variability are in one group while low ones are in the other. Murphy's law virtually insures that this will happen.

The term "Alpha" is used to indicate the probability of making a Type 1 error. For example, if you repeated an experiment 100 times with virtually homogeneous subjects drawn from the same population, you could expect some random variation in the results just due to chance. A probability of 0.05 means that five times out of a hundred (one time out of twenty), the groups in your study are as different from each other as you found by chance alone. So, you have a one in twenty (five percent) chance of making a Type I error if you say the groups are different from each other. Most clinicians do use a 0.05 level of significance as they consider this to be ample protection. Problems with this decision are discussed further in the statistics section.

The term "Beta" is used to indicate the probability of making a Type II error (saying that there is no relationship when there is one) . People are usually satisfied with at Beta of 0.20 which means that you have a twenty percent chance of making a mistake. The power of a test is 1-Beta so you would set your test's power at 0.80 if you set the probability of a Type II error at 0.20. As you decrease Alpha and Beta, the number of subjects required to meet your criteria increase exponentially so you have to trade-off how certain you wish to be of being correct against the number of subjects you can afford to run.

3. Level of confidence that you will achieve a 0.05 (or whatever your choice) level of significance: We usually work with 95% confidence intervals / levels. The idea of confidence intervals will be discussed further in the section on regression. It is the idea that 95% of your measures will be within plus or minus one standard deviation of the number the confidence interval is around. Setting your confidence level at 95% means that you have a 95% chance of actually finding the groups different at the level of significance you set.

4. Effect size - What amount of difference is important? As noted above, sample size determinations count of balancing variability against amount of change to determine how many subjects will be required. The investigator has to supply the information about how much change is important. This is the investigator's chance to avoid having a difference declared "statistically significantly different" when the actual amount of difference is clinically meaningless.

C. Formulae for determining group size: There are many approaches to determining group size. They vary with the structure of your study, the statistical test(s) you intend to perform, and the type of data you will gather. Whole books are written on this subject.

1. General formulae: Ostle (1963) published a reasonable formula for determining rough "ball park" group sizes, regardless of statistical test or type of data. It should only be used when more specific formulae are not available as the results could be quite misleading. The following abstract from the tables is based on using a 0.05 level of significance with a 95% confidence level.

In order to use the formula, you need to provide (1) a reasonable, **clinically significant**, change from normal / control values (or change from baseline or difference between groups) and

(2) the standard deviation of the sample data (the amount of variability you can expect around the mean value for your normal group).

Start by dividing (1) above by (2) above to find the power factor.

$$\text{FACTOR} = \frac{\text{(1) difference between means}}{\text{(2) standard deviation}}$$

Use Table 15 to relate the factor to group size.

Table 15 Group size factors

Factor	Group Size	Factor	Group Size
2.4	6	1.5	13
2.2	7	1.4	15
2.1	8	1.3	17
1.9	9	1.2	20
1.8	10	1.1	23
1.7	11	1.0	27
1.6	12	0.95	30

If you were working with diabetics and wanted to try several interventions to see which was more effective, you might decide that a mean difference of 20 in blood sugar was the smallest difference which would be clinically important, you would look at the blood sugar levels of the type of patients likely to be included in your sample by reviewing records and etc. and then calculate the standard deviation of the readings. Let's say that the standard deviation came out to fifteen. Dividing 20 by 15 results in a factor of 1.3 so 17 subjects would be needed in each group to be 95% confident that you would find a difference at a probability of 0.05. However, if the standard deviation was only ten, the factor would be two so you would only need eight subjects per group. The same effect would be achieved by increasing the predicted amount of difference between the groups. Unfortunately, when people actually apply these formulae, they tend to redo the test several times while adjusting the various factors so they eventually find a combination which gives them a number of subjects they can afford regardless of whether a bit of reality has to be bent to get there.

2. Formulae specific to the statistical test to be applied:

a. These are complex and found in most statistical programs such as “primer” and “SPSSPC.” There is no need for anyone working at the basic clinical level to attempt doing the calculations by hand.

b. I am not aware of any sample size tests for non-parametric statistics, logistic regressions, or categorical variables so you need to use the most similar parametric test which would be used if the data met the entrance criteria for the test.

c. For descriptive studies, base your power analysis on the interval confidence level.

d. If both the predictor and outcome variables are dichotomous (only two choices - yes/no), use the “z” statistic (Chi Square). If one is continuous (e.g. weight) and the other is dichotomous, use a “t” test even if this is a non-parametric variable (e.g. pain levels). If both are continuous, use a correlation coefficient.

e. Examples of results from computer based formulae:
All of the following had alpha equal to 0.05 and power equal to 0.80.

(1) Paired “t” test with the difference between the means equal to 20:

If SD (standard deviation) = 5, then N/group = 3.

If SD = 10, then N/group = 4.

(2) Unpaired “t” test with the difference between the means equal to 20:

If SD = 5, then N/group = 4.

If SD = 10, then N/group = 6.

(Note that less subjects are needed for a paired “t” test than for an unpaired one because the variability is not due to differences between subjects but, rather, to something that happened between the first and second administrations of the test.)

(3) ANOVA (parametric, one way, non-repeated measures analysis of variance) for three groups with the difference between the means equal to 20:

If SD = 5, then N/group = 3.

If SD = 10, then N/group = 6.

(4) Proportions of two groups:

If the proportion of group one showing the problem is 0.6 (60%) and that of the second is 0.3, then N/group = 43.

If the proportion of group one showing the problem is 0.9 (90%) and that of the second is 0.2, then N/group = 7.

(5) Correlations between two groups (linear, parametric):

If the anticipated correlation is 0.8, then N/group = 10.

If the anticipated correlation is 0.4, then N/group = 47.

(6) Chi square (2X2) contingency:

If the expected ratios are 2:1 vs. 1:2, then N/group = 68.

If the expected ratios are 2:1 vs. 1:3, then N/group = 45.

If the expected ratios are 1:2 vs. 1:5, then $N/\text{group} = 209$.

D. Sample size for surveys: Sample size techniques need to know how confident you want to be that your sample will be representative of the population and either the proportion of the population likely to respond in a certain way to a key question or how variable the responses are likely to be. For example, if you want to know what proportion of war related amputees have shocking phantom pain and the Veterans Administration is willing to supply you with a list of addresses for all 25,704 US military veterans known to be amputees, you would begin by making an educated guess from your experience and the literature of what the anticipated rate is. Let's say that the literature says that the rate is between 10 and 20 percent but you feel that it is closer to 20 percent. Use the smaller number so you do not survey too few people. The next question is what margin of error you will accept in your certainty that your sample represents the population. Most people use 95% certainty (a five percent margin of error) but I tend to use 98% because I have seen too many flukes happen. The usual formula (Bordens and Abbot, 1991) is: sample size equals the proportion of the people you feel will have the problem (.1 in our example) times 1 - the proportion (1-.1) divided by the square of the margin of error (.02). So,

$$n = (0.1)(1 - 0.1)/(0.02)^2 = 0.09/0.0004 = 225$$

If you have such a small population that the sample size is ten percent or more of your population, you need to adjust your sample size so you do not over sample. The formula (Bordens and Abbot 1991) is: adjusted sample size equals population size times the original sample size divided by the population size plus the original sample size. For example, if your population was the 1,400 amputees seen by your hospital in the last three years, the calculation would be:

$$\text{adjusted } n = (1,400 \times 225)/(1,400 + 225) = 315,000/1,625 = 194$$

Do you seriously believe that only 225 surveys will tell you what you want to know? I don't.

Very often your survey is attempting to find out if your intervention worked better than another intervention. This means that you need to find out how many responses will be needed to tell the difference between two proportions. Simply use a standard power analysis program to determine how many people you will need. For example, if you feel that 20% of the patients get better with the standard technique but 60% get better with yours and you want to be 95% certain that your groups will differ by your projected amount (a 0.05 level of significance) with a 15% chance of concluding that the groups are not different when they really are (a power of 85%), you would only need 31 people per group because of the huge difference in expected outcomes (40%). However if you felt that half the people got better with the standard technique then you would need 463 people per group because there is only a ten percent difference between the two groups.

E. What if there are no data to use for estimating sample size? Unfortunately, (1) there is frequently no trial data available upon which estimates of variability, and thus, of number of subjects, can be based and (2) it is unfeasible due to constraints on time, investigator availability,

and other aspects of real military life to run a separate pilot study. In this case, the accepted method of establishing the appropriate sample size is to estimate the maximum number of subjects likely to be required (based on background knowledge of the population and the disorder under investigation) and to request approval to use that number of subjects in the study. When the study is proposed, the minimum number of subjects likely to show differences in response is estimated from background knowledge. The usually accepted MINIMUM number of subjects for this initial estimation for clinical studies is five per group when two or more groups are to be compared or ten when the group is to be compared either with itself (pre - post measures) or with the general population. The data on the most important variables should be evaluated as soon as this minimal number of subjects complete participation in the initial data gathering stages of the protocol. If it becomes evident that the techniques cannot determine any difference between the groups (or etc.) , no further subjects should be entered into the study. If it is already statistically and clinically evident that the groups (or etc.) are different, there is also no further need to proceed with the study. If there are differences between the groups, but the differences are not sufficiently great to assure the investigators that the differences are due to more than those which could occur from random sampling of small groups of humans (with their inherently great variability), power analysis is used to estimate how many more subjects should participate to determine whether there is a REAL, reliable, replicatable, clinically important difference between the groups (or etc.). If this number is greater than the maximum number of subjects initially approved for use in the study, the investigators have to demonstrate to the approving committee that it is feasible to complete the study and that sufficient subjects are available.

The following is a sample of how to phrase the section for group size when you really have almost no idea of how many subjects you will need:

"We do not have sufficiently detailed data on individual response variability from previous studies to predict the actual number of subjects which will be necessary to differentiate our response rate from theirs. We can get about 10 - 15 patients per year and can continue to the study for two years. As the comparison studies used 30 and 11 subjects respectively, thirty is a reasonable number of subjects for comparison. Thus, we request permission to use up to 30 subjects in the study. After the first ten subjects are completed, we will analyze our data and compare our response rates with those of the other studies using Chi Square proportion techniques. If there are no differences or if the groups are already significantly different at the $p \Rightarrow 0.05$ level, we will terminate the study at this point. If there are trends toward differences between response rates, we will use power analytic techniques to determine the number of extra subjects likely to have to be included to give us an 85% chance of differentiating between the results of our study and the comparison studies at the above probability level. If the predicted number exceeds 30 subjects, and we feel that we have sufficient time to enter the requisite number of subjects, we will request the committee's approval for expanding the size of the study pool."

F. Establishing safety: One of the aims of clinical studies is frequently to assess the safety of an intervention. If the population studied shows a particular adverse reaction one percent of the time (e.g., likelihood of getting a fever during your study), according to Spiler (1986), you would

need 300 subjects to have a 95% chance of observing one adverse reaction during the study. At a more usual rate of one adverse event in 1,000, you would need 3,000 subjects to detect it. For a relatively rare event such as 1/50,000, you would need 150,00 subjects so your twenty subject study simply does not cover “safety”. If you truly wish to establish that a technique is safe, you will have to perform very large studies at relatively high dosages in which your subjects are followed for a very long time.

Sample Study

Look at the subject selection sections of the first two studies.

Are the methods likely to lead to a random selection of typical migraine headache patients?

Are the inclusion and exclusion criteria clear and reasonable?

Chapter 28

Evaluating overlap between groups using Inferential statistics

A. Concept: When we look at scores on some test generated by two groups or by one set of patients measured before and after treatment, we generally find that the scores of the two groups or periods overlap. We normally want to know whether the two groups of scores are "really" different from each other or just look as different as they do by "chance." In order to use statistics to support (*not make*) a judgement about the likelihood of the groups being as different as they are by chance, a basic understanding of the tests is required.

Imaginary data which could have been generated by the sample controlled study are provided on the next page for numeric examples. It must be noted that, in an actual analysis, the sample tests would not be performed on these data in the order they are done and much of the analysis would be done using tests of response frequency rather than intergroup tests for reasons which will become apparent as you learn about the tests and their uses.

B. Tests for differentiating between two groups:

1. There are four common tests (Paired "t" test, Independent "t" test, Wilcoxon Rank-Sum, and Mann-Whitney U). Which is used depends on the nature of the data and the study design. If you tested the same subjects twice, most of the changes in the numbers should be due to factors such as your intervention and time between the tests. Since you have "repeated measures" from each subject, you can use "paired" data tests. If your data are parametric, you can use a paired t test as long as your data meet the other assumptions for the test (most importantly that the two groups have very similar variability and that both distributions are normal). If your data are non-parametric, you can use a Wilcoxon Rank-Sum test. If you are not working with repeated measures, much of the difference between the groups may be due to inherent, random differences between the subjects participating in the groups. Thus, a more conservative test must be used. For parametric data which meet the entrance requirements, a Student's independent t test is used (as opposed to the paired t) while a Mann-Whitney U test is used for non-parametric data.

	repeated measures	independent measures
parametric data	Paired "t"	Independent "t"
non-parametric data	Wilcoxon Rank Sum	Mann-Whitney U

Sample study - Imaginary data which could be from the controlled sample study

G R O U P	SUBJECT #	Average headache activity per week						NUMBER YEARS OF MIGRAINE
		1 MONTH BASELINE		1 MONTH POST- EXPOSURE		1 MONTH FOLLOW- UP		
		freq	inten	freq	inten	freq	inten	
Actual Ex- posure	A1 (aura)	6	9	2	3	1	1	43
	A2 (aura)	6	7	2	2	0	0	12
	A3 (aura)	5	8	3	8	5	8	19
	A4 (aura)	7	6	5	4	3	1	31
	A5 (aura)	8	9	4	0	0	0	40
	A6 (no aura)	5	5	4	4	1	2	23
	A7 (no aura)	4	7	4	3	2	1	30
	A8 (no aura)	9	8	7	8	9	7	25
	A9 (no aura)	7	6	5	6	5	4	37
	A10 (no aura)	6	8	4	5	4	3	22
Placebo Ex- posure	P1	8	7	7	7	8	7	31
	P2	6	6	7	3	5	5	24
	P3	7	9	6	8	6	9	16
	P4	9	8	9	5	6	8	34
	P5	5	5	6	6	5	5	17
	P6	6	7	5	6	6	6	19
	P7	4	9	3	9	1	1	22
	P8	8	9	7	6	8	8	28
	P9	6	6	5	6	5	6	18
	P10	7	8	5	7	7	7	33

2. The Independent Student's and Paired "t" tests:

Use: For testing differences between two groups of continuous variables.

Assumptions for use of the "t" test: Both groups are normally distributed with similar variation, random samples from populations with similar variability. The test is NOT for testing between samples chosen with stratified sequential sampling because the samples are not random.

Sample size for the "t": minimum of 3 to infinity.

Degrees of freedom (DF) for "t" = $n_x + n_y - 2$

Significance for "t": Printed out by your computer. If you are looking up "t" values in a book, you are doing extra work because it is just as much work to enter the values into a calculator as into a computer which can do far more for you.

Sample study data - The independent "t" test

"Years of migraine" is a continuous variable. Determination of whether any differences in number of years of migraine between the actual exposure and placebo groups are statistically significant can be tested using an independent "t" test as long as the variances are similar and the two groups are normally distributed.

Your program should print out at least the group means and standard deviations.

When a standard deviation is shown in a summary, it is abbreviated as "SD" or as "+/-".

A sophisticated program will tell you whether the distributions are variances are sufficiently similar for the test to be valid. In this case:

Actual exposure: mean = 28.2 +/- 9.8

Placebo exposure: 24.2 +/- 6.9

"t" = 1.06 with 18 DF; $p = 0.31$. Thus, the number of years of migraine does not differ significantly between the actual and placebo exposure groups.

3. Mann-Whitney U and Wilcoxon Rank Sum:

Use: For testing differences between two non-parametric samples or between a sample and its population.

Assumptions: Random sampling

Significance: Printed out by your computer as discussed above.

Sample study data - The paired “t” test

The original study design does not include a follow-up period. Thus, if you were performing an analysis of the data before the follow-ups were done, you would have only one before treatment and one after treatment measure to consider. This means using “t” tests for the parametric measures and either the Wilcoxon or Mann-Whitney U for the non-parametric data. “Change in frequency of headaches” is a parametric variable. If you wanted to determine whether frequency changed from baseline to 1 month post exposure for the actual exposure group, you would use a paired (repeated measures) “t” test.

“t” = 5.44 with 9 DF; $p = 0.0001$. Thus, the frequency of headaches changed statistically significantly between the baseline and post-treatment periods.

Sample study data - The Mann-Whitney U and the Wilcoxon Rank Sum tests

Pain intensity is a non-parametric variable so changes in pain between groups need to be evaluated using the Mann - Whitney U test while changes between periods within one group need to be evaluated using the Wilcoxon Rank Sum test.

The difference in pain intensity between baseline and the post-treatment month for the actual exposure group is a repeated measure as the same people were evaluated twice. The Wilcoxon produced a “W” of 28 which shows a significance of $p < 0.06$ so pain did not change statistically significantly during this period.

The difference in pain intensity between the placebo and actual exposure groups during the baseline is an independent groups measure so the Mann-Whitney U test is used. Ten subjects per group was enough for the computer algorithm to determine that a “t” statistic could be used. “t” = 0.154 with $p = 0.88$ so the groups did not differ statistically.

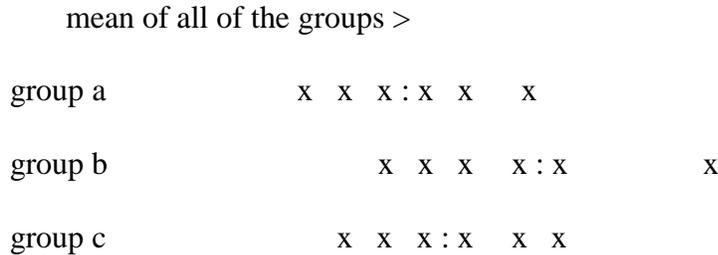
C. Tests for differences between more than two groups - Analysis of Variance (ANOVA):

These tests are fairly similar to “t” and “u” tests and have similar restrictions in their use. When a one way, parametric ANOVA is applied to 2 groups, it gives the same answer as a “t” test.

1. Concept: ANOVAs work by combining the variability and scores of all of the groups to get a “grand” mean and standard deviation. If the grand mean = 100 and the grand standard deviation equals 5, then (when reasonably sized groups are recorded), about 66% of the sample groups will have means between 95 and 105. This is because 66% of the population is expected to be within one SD of the mean. 95% of the means would be within 1.96 SDs of the population mean or between 90.2 and 109.8. Thus, the odds of a mean being outside this area by chance can be figured out. Figure 12 shows the relationship between group means and the grand mean.

Figure 12 Relationship of groups mean to the grand mean for all groups

(The : symbol indicates group means; the vertical line indicates the grand mean of all groups combined. x = individual subject's score).



2. Assumptions of parametric ANOVAs: (a) subgroups normally distributed, (b) random sampling, and (c) similar variance for all groups

3. Problems in the use of ANOVAs with clinical data: Parametric ANOVAs use the amount of variation in all of the groups to estimate the "population's" variance. Thus, if the test causes a change in variability, a parametric ANOVA can not be used.

4. Non-parametric ANOVAs: Similar to parametric ones but use overlap of ranks and produce a chi square statistic rather than an "f" value. Use these when your data is non-parametric or when group variability is not similar. The Kruskal-Wallis test is used for independent measures while the Friedman test is used for repeated measures (e.g. when patients are measured several times).

5. Structure of ANOVAs: ANOVAs can be one or two way and can be set to look for repeated or independent measures. A "one way" ANOVA would be one where you are comparing one reading from several groups (independent) or several readings made on one group (repeated measures). A "two way" ANOVA is one where you are comparing two different sets of changes at once. For example, you might compare three intensities of three different treatments. This would require you to have three treatment groups to compare (one of the two ways) and three doses for each treatment (the second of the two ways). The design is illustrated in Table 16. When your observations are repeated several times on the same subjects, use a repeated measures analysis of variance. If you have several groups and each has repeated measures, use a two way, repeated measures analysis of variance which is set-up for independent groups with repeated observations. If the design shown in Table 16 used repeated observations of each group rather than different doses, it would be a two way, repeated measures ANOVA.

Table 16 Structure of a two way analysis of variance

	Treatment 1	Treatment 2	Treatment 3
Dose 1	group 1 (receives Rx 1 and dose 1)	group 4 (receives Rx 2 and dose 1)	group 7 (receives Rx 3 and dose 1)
Dose 2	group 2 (receives Rx 1 and dose 2)	group 5 (receives Rx 2 and dose 2)	group 8 (receives Rx 3 and dose 2)
Dose 3	group 3 (receives Rx 1 and dose 3)	group 6 (receives Rx 2 and dose 3)	group 9 (receives Rx 3 and dose 3)

Sample study - Parametric and non parametric one way analysis of variances

1. Non-parametric: To evaluate statistical changes in intensity of headaches over the three observation periods (baseline, 1 month, and follow-up) for the actual exposure classic migraine group, a one way, repeated measures, non-parametric ANOVA needs to be used. The Friedman test produces a Chi Square output of 5.7 which is statistically different at $p = 0.005$. Note that you can't tell which groups differ without doing Wilcoxon tests between each group and then correcting the level of significance for over-testing using a technique such as the Bonferroni test discussed below.

If you were comparing three independent groups (perhaps actual exposure classic, actual exposure common, and placebo), you would use the Kruskal-Wallis test.

2. Parametric: To make the same comparison as above but for frequency of headaches, you would use a parametric, one-way analysis of variance. The "F" = 11.27 which indicates a statistically significant difference as $p = 0.005$.

To compare frequency across the three groups suggested above, you would use a parametric, one way ANOVA.

6. Two way analysis of variance and interaction effects: Often several factors will vary in the same study. For example, a study might examine which of three dosages of a drug (e.g. low, medium, and high) are most effective for controlling a condition and simultaneously examine the best way to distribute the dose throughout the day (e.g., give it all during the morning, spread it evenly throughout the 24 hour day, or give it all in one shot at night) to optimize effectiveness. The effects of these two manipulations may not be independent of each other. A low dose given at night may be more effective than a high dose given in the morning. The interaction of the variables is called the "interaction effect." When you perform a two way ANOVA, always look

for this effect. Many people ignore it because it means that the study's results are more complex than predicted (and, thus, more complex to understand and to report).

7. Factoring out known confounding variables: If you did a study in which the observations are affected by time of day and you were unable to control the time observations were made, but you did know the time each was made, you can use special formulae to determine the impact of time on each observation and then alter or weight the numbers used in the analysis to compensate for the effect of the time each observation was made. This analysis is called an analysis of co-variance or an ANCOVA.

D. Correcting for performing multiple two group tests: You must use an ANOVA prior to performing multiple "t" or other two group tests even when the results of the multiple "t"s are corrected using special formulas such as the "Bonferroni correction for multiple comparisons". This is done because (1) as explained above, if you have a 5% chance of making a mistake and run 20 combinations, you should get two of your groups showing as different by chance alone, (2) multiple tests can not use the variability from all of your groups at once so tend to indicate that two groups are statistically different when their means simply happened to be relatively further apart than the means of the other groups by pure chance, and (3) the correction formulae are frequently too conservative for clinical tests so they will indicate the lack of a difference when there probably is one. When using the correction, divide the "p" value used by the number of tests to get the corrected "p" value. For example, if you want to use a "p" of 0.05 and you do four tests, $0.05/4 = 0.0125$ so your corrected p value is now 0.0125 which gives you the same level of assurance you won't make a Type I error as 0.05 would have given for one test.

E. Meta-analysis: Very frequently, numerous studies have already been performed in an area of interest to you. For example, hundreds of studies have used drugs to treat migraine headaches and dozens of studies have used behavioral interventions for the same purpose. If you wanted to compare the effectiveness of behavioral and medicinal interventions for migraines, you could perform a huge, expensive study or you could try to compare the outcomes from the two sets of studies already performed. Comparing any two studies is a very real problem because they rarely, if ever, use the same population, techniques, etc. Even when one study purposely attempts to exactly replicate a prior study, the results rarely turn out very similar to each other. One approach is to perform a very good review of the literature and make tables comparing the outcomes of the different studies as best as you can identify them in the results sections. These lists tend to be long and difficult to interpret as each study has a different design, different numbers of subjects, and different levels of change reported. Few of these types of reviews reach conclusions which are acceptable to most readers as being firm enough to warrant changing clinical practice.

Meta-analysis is a group of statistical procedures which make it possible to combine the results from many studies to overcome the above weaknesses. These techniques mathematically combine information about number of subjects, effect size and level of significance from each acceptable study into a conglomerate outcome. The trick is to find enough studies which (a) provide sufficient key information, (b) are similar enough in subjects and design to make valid

use of their data, and (c) are of high enough quality so the results are trustable. What usually happens is that the reviewer begins with dozens of articles which supposedly cover the topic and winds up with only a few which meet the requirements for combination. The results can be important because combining ten tiny but well done studies with ten subjects each can give the results for a hundred subjects treated in a similar way. The idea is that variability from each group can be combined to give a better idea of overall subject variability across studies in the same way an ANOVA combines the variability between groups. Smith et al (2000) used acupuncture for chronic neck and back pain as an example. They found that meta-analyses which didn't eliminate poor quality studies concluded that treatments were far more effective than was the case – as demonstrated by meta-analyses which did eliminate the poorly done studies with untrustworthy data.

Meta-analyses are usually biased because studies that don't show differences between groups tend not to be published and they are relatively weak because most publications don't provide sufficient data for meaningful use to be made of their work.

Bordens and Abbott (1991) provide a clear description of how to actually perform the mathematics of a meta-analysis. I recommend you hesitate a long time before you attempt one by yourself because making a valid meta-analysis has defeated many of the more sophisticated clinicians who have tried it on their own. Frankly, study designs and subjects tend to be so different and data presentations so varied that little of real value comes out of most of these attempts. I strongly urge you to take the results of any meta-analysis with a large grain of salt and to draw your own conclusions from the summary of the data presented.

Chapter 29

Evaluation of relationships between changing variables

A. Correlation:

1. Correlation is a measure of how one variable changes in relation to changes in another or others. Correlation *does not indicate* that a change in one causes a change in the other!!! Association does not equal cause. For example, there is a high correlation between foot size and spelling ability in elementary schools. Correlations range from zero to either plus or minus one. Sign does NOT effect the significance of the relationship. The size (magnitude) of the number "r" indicates the extent, but not the "statistical or clinical significance", of the relationship. Plus 1.0 or minus 1.0 means a perfect correlation while 0.0 means no correlation at all. A correlation would be perfect if the two variables changed in lock step with each other while it would be zero if the changed totally randomly with respect to each other. The correlation coefficient is indicated by "r" while "r²" indicates how much of the variability is accounted for by the correlation.

Statistical significance is a combination of the number of pairs and the magnitude of the correlation. Just as is true for tests between groups, there are parametric (Pearson) and non-parametric (Spearman) correlations. Also, as is true for tests between groups, multiple correlation tests are used when there are more than two groups rather than performing multiple "single" correlations. Figure 13 shows positive and negative linear correlations and Figure 14 shows the relationship between how closely the variables change with respect to each other (tightness of fit) and the strength of the correlation.

Figure 13 Positive and negative linear correlations

Positive correlation (+): both variables change in the same direction.

Negative correlation (-): When one variable goes up, the other goes down.

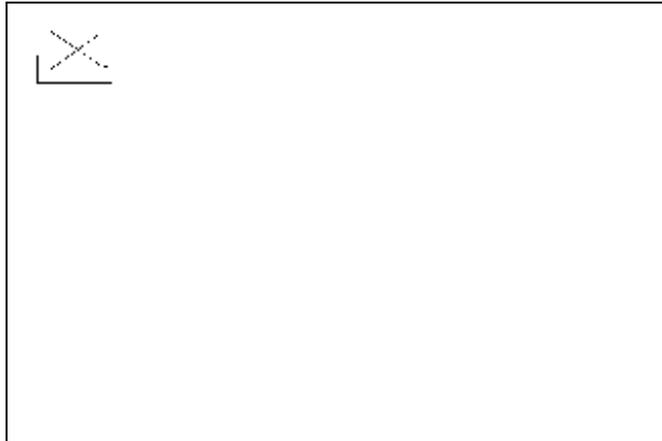
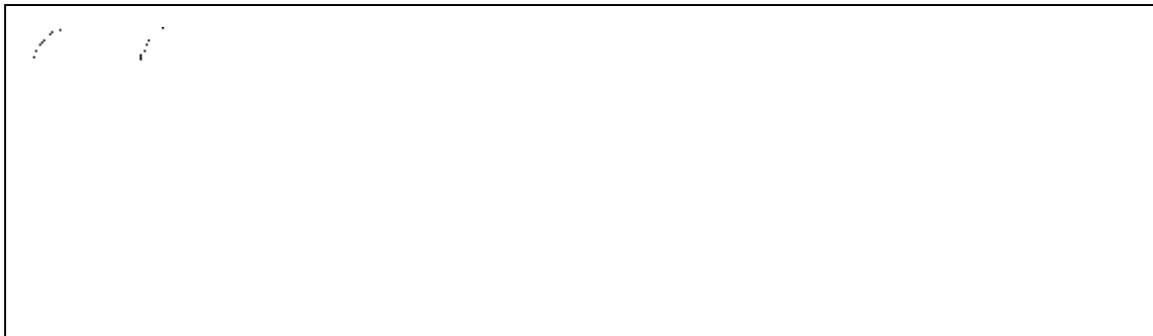


Figure 14

***Relationship between tightness of fit to the correlation line
and strength of a curvilinear correlation***

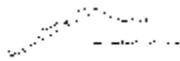
*"Good" correlation -
tight fit of points
to line*

*"Poor" correlation
loose fit of points
to line*



2. Nonlinearity: Life is nonlinear rather than linear and even linear correlations are usually not perfect in the clinical environment because each patient responds a little differently. Statistical packages draw a "best fit" line to match the data as closely as possible. This is usually done with the "least squares" technique in which the distance between the line and all of the points is minimized. If the line curves, the relationship is NOT linear. Variables do not have to vary together arithmetically (which would produce a linear relationship) to be highly correlated. They may have many different relationships. The important thing is how consistent the relationship remains as the variables change. Figure 15 illustrates producing a "best" fit line from slightly non-linear data. If the data are highly non-linear, either a non-parametric correlation is used or the data are transformed to make the relationship linear.

Figure 15: Producing a "best fit line from slightly non-linear data



3. The major two-group correlational tests:
 - a. Pearson's "r" (parametric / linear) uses actual numbers and means. When used with non-linear data, it greatly underestimates correct amount of correlation so you must change the data before using this test.
 - b. Spearman's "r" (non-parametric) uses ranks and differences between ranks of pairs. It does somewhat better than Pearson's but is still not great at estimating correlations of curved lines.

Sample study - linear / parametric correlation using Pearson's "r"

When looking at your study population, you may want to know if frequency of migraines is related to years of migraines because people who have had headaches may be more difficult to treat. As both frequency and years are parametric measures, Pearson's correlation can be used. It produces a correlation coefficient of 0.275 with a "t" of 0.81 (8 DF) and a probability equal to 0.441. Thus the two variables do not change with each other.

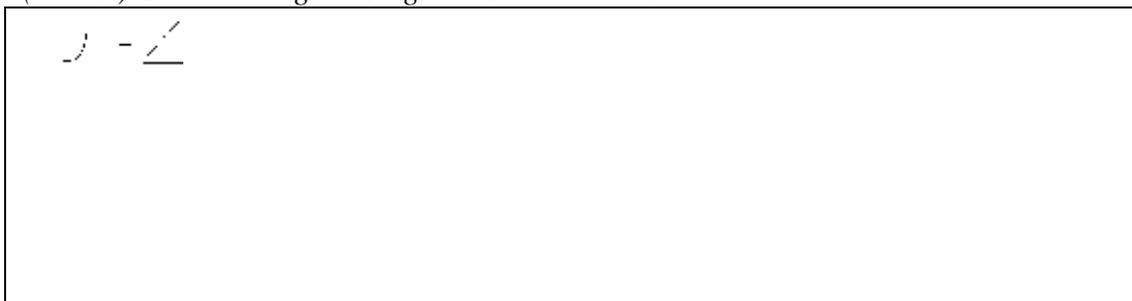
A good statistics program should graph out your data so you can get an idea whether the relationship is linear. If it isn't, the test's results may be meaningless.

4. Handling nonlinearities by transformation of the data to straighten curves: This is done so you can apply the above statistical techniques to determine the strength of the relationships. These techniques can be done automatically by your computer. NOTE: Do not do this without looking at **both** the raw data scatter plot and the best fit curve or you will miss the effect that outliers are having on the curve produced. Once the data are transformed, you can't tell what the actual relationship looked like! What if the data hug the curve closely for one stretch and then clearly change in relationship so they hug it loosely at an other segment? Obviously something has happened! Use different correlations for the two areas and try to figure out what happened at the break point. This may be your most important result.

Figure 16 shows several common ways of transforming non-linear data so powerful linear correlation techniques can be used to evaluate relationships between the variables. The transformations were developed because the curves shown in the examples occur very often in clinical work.

Figure 16 Common methods of transforming non-linear data

16a. Semilog transformation changes smooth single bend curves into straight lines (almost). It also changes straight lines into curves.



16b. Reciprocal of number is used for hyperbolic curves.



16c. Normal Equivalent Deviates and PROBITS are used to straighten "S" shapes.



B. Prediction with Regression:

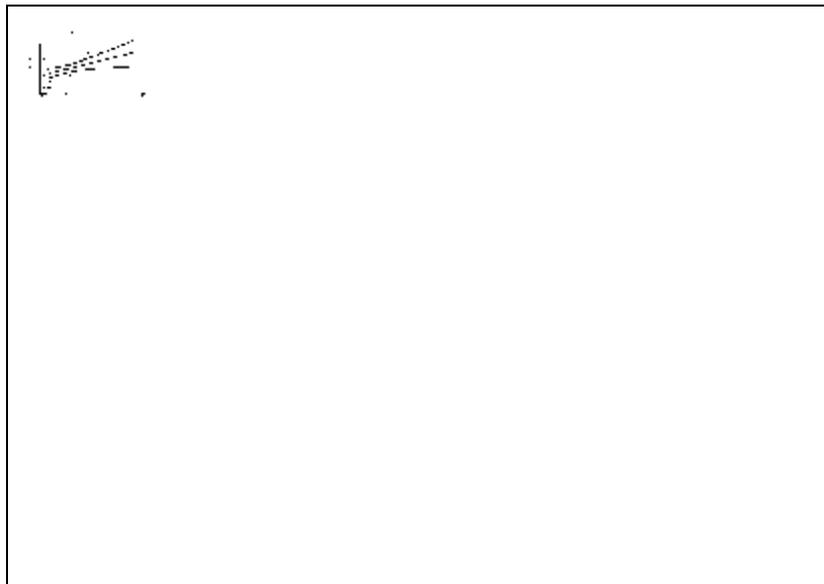
1. If you know the correlation and "best fit" curve for the population, you should be able to draw a line from a known number on one axis to the curve and then to the other axis to find out what the other member of the pair is likely to be. This idea is illustrated in Figure 17.

Figure 17: Regression from one variable to another



2. The accuracy with which you can predict the value of one number of the pair from the other depends on how tightly the regression line fits your actual data. If you don't have a very good correlation so that there is lots of "slop" around the line, then you can't be particularly certain of the value of the other member of the pair. If the correlation is very good, you can be relatively more certain of the other member's value. The statistical method for evaluating how certain you are of the value is called the "confidence limits of the regression line." They are based on "t" scores and are interpreted the same way. Confidence limits are expressed as numbers falling on either side of the pair's predicted value. When you use 95% confidence limits, you are predicting a result with 95% accuracy. Thus, if you predict a value for "y" of 9 with 95% confidence limits of 2, then 95% of your values should occur between 7 and 11. Figure 18 illustrates the 95 percent confidence limits around a linear correlation line of best fit.

Figure 18 **Confidence limits around a correlation line**



3. Multiple regression techniques: Many variables are frequently correlated with each other. Multiple regression techniques can tell you which combination of variables give the best ability to predict. They give you the percent of the variability each variable accounts for in the correlation. The investigator has to decide whether the contribution of a particular variable is enough to say it is a valuable predictor of the overall relationship.

C. Comparing two correlations: It is frequently very important to know whether two variables change at the same rate. For instance, two dose - response curves may reach the same level eventually but one may appear to get there faster. This could indicate that a treatment works

more quickly. When you have two correlation lines to compare, use programs designed to look for differences in the lines' slopes. Programs to compare differences between the slopes of two curves are not readily available so you have to transform curved lines into straight ones as best as possible using the techniques discussed above. An alternative approach is to look at the most meaningful parts of the lines and only transform those parts into straight lines. This can eliminate unnecessary distortion of the data and, thus, produce a more realistic result.

The left side of Figure 19 illustrates a pair of typical dose-response curves which are typical of responses to clinical interventions and which need to be transformed into the straight lines shown on the right side of the figure using Normal Equivalent Deviate and PROBIT techniques.

Figure 19 Straightening a pair of clinical intervention dose-response curves using Normal Equivalent Deviate and PROBIT techniques



Unfortunately, in the real clinical environment, two dose response curves rarely have the same shape so the same formula can not be applied to straighten both to the same extent. This makes for a very sticky statistical situation as the transformed data for one line will fit the correlation formula better than the other. In such cases it is better to use the suggestion made above to just look at the most important parts of the curves so extraneous data are minimized.

Twelve weeks of typical dose-response data for people with classic (with auras) and common migraines (without auras) exposed to pulsing electromagnetic fields are shown in Table 17 and illustrated in Figure 20. Note that the solid line representing the responses of subjects with classic migraines flattens out at the eighth week. If the lines are not transformed to minimize distortion of the data, the flat right tail must be cut off before the data are analyzed or the formula will compute the wrong slope for the portion of the line with interest. With all 13 pairs, the slope of the line is -0.55 (standard error = 0.05) but, with only the relevant nine pairs, it is -0.70 (SE = 0.04). The correlation with all pairs is 0.96 (t = 11.3 with 11 DF) while that with nine pairs is 0.99 (t = 19.9 with 8 DF). Thus, it makes a very real difference if the wrong parts of the line are included. The slope for the common headache responses is -0.50 with a standard error of 0.04. If the wrong part of the data (all 13 pairs included) were used, the statistical

software would think the lines had very similar slopes.

A “t” test of the difference between the slopes ($t = (\text{slope one minus slope two}) / \text{square root of (standard error 1 squared plus standard error two squared)}$) is 3.51 with 18 degrees of freedom (total number of pairs minus four) which is significant with a probability of at least 0.005.

Table 17 **Twelve weeks of dose - response data for subjects with common and classic migraines exposed to pulsing electromagnetic fields.**

Type of migraine	Frequency of headaches per week after “x” weeks of exposure												
	0	1	2	3	4	5	6	7	8	9	10	11	12
classic (with aura)	6	5	5	4	3	2	2	1	0	0	0	0	0
without aura	6	6	6	5	5	5	4	4	3	2	2	1	0

Figure 20 **Dose - response curves for subjects with common and classic migraines exposed to pulsing electromagnetic fields for twelve weeks.**



D. Time series analysis - a special type of correlation:

Very often you want to determine not only how two variables change with respect to each other, but how that relationship changes over time. You may also want to determine whether time has some recurring effect on changes in your outcome variable. For example, if you were evaluating reports of migraine headache activity in women, you would want to know if their headaches came periodically - with the menstrual cycle or other periodic events such as monthly paydays. In order to evaluate periodicity, you would perform a time series analysis which looks for patterns in changes over time.

Chapter 30

Dichotomous and proportional data

A. Use of clinical test results in research: The results of standard clinical tests are commonly used as outcome variables in research. Diagnostic tests are designed to detect the presence of a problem while prognostic tests are intended to determine the outcome of the disease. The test result is a predictor variable which the investigator hopes is very highly related to the actual outcome variable of interest which is presence or absence of the disease.

Most clinical tests do not give a yes/no answer, but rather some representation of severity. This means that a cut off point must be established beyond which the test's answer is highly likely to be indicative of the disease being present. Unless the test is 100% accurate, establishing the cut off point requires making a trade off between the test's sensitivity (correctly determining that the patient has the disease) and specificity (correctly determining that the patient is not diseased). The cut off point has to be set depending on whether it is more important to catch all of those with the problem or miss a few. If the disease is very difficult to detect at a stage when treatment can still be effective and kills off just about everybody who gets beyond that stage, the cut off has to be set very low so you don't miss anybody. This means that you will have many "false positives" but very, very few "false negatives".

The data from tests such as these can not be used as outcome variables unless you know how efficiently the test is performing. Techniques for "testing" the efficiency of outcome measures are reasonably well worked out.

B. Evaluating how good (efficient) a test is: Many clinical laboratory, diagnostic, and efficacy tests give dichotomous (yes/no, sick/healthy, normal/abnormal) results. For example, when evaluating a series of muscles, the clinician usually wants to know if each muscle or muscle group falls into the normal range of activity. The degree of activity (level of the outcome parameter) is only of interest if it is abnormal. Unfortunately, few (if any) tests are perfectly accurate.

An ideal test would only give a positive result when the person was actually sick (a true positive) and a negative result when the person was actually healthy (a true negative). However, tests aren't perfect so they miss in both directions at least to some extent by giving false negatives (the person is actually sick but the test fails to show it) and false positives (the person is actually healthy but the test declares for sickness). Before using any test, you need to know how accurate it is. Accuracy is determined by the test's predictive value and efficiency. In the following definitions, TP = true positive, FP = false positive, TN = true negative, and FN = false negative. Remember that when discussing probabilities, being absolutely certain is a probability of 1.0.

The false positive rate is the probability of incorrectly classifying a healthy person as sick.
The FP rate = (number of false positives)/(total number of healthy people tested).

The false negative rate is the probability of incorrectly classifying a sick person as healthy.
The FN rate = (number of false negatives)/(total number of sick people tested).

Sensitivity is the *probability of correctly classifying a sick person as sick*. A test's sensitivity is the frequency of positive results in patients who really have the problem.
% Sensitivity = $(TP/(TP+FN))*100$ = (number of true positives)/(number of sick people tested) X 100.

Specificity is the *probability of correctly classifying a healthy patient as healthy*. A test's specificity is the frequency of negative results in patients who do not have the problem.
% Specificity = $TN/(TN+FP)*100$ = (number of true negatives)/(number of healthy people tested) X 100.

Predictive value is the frequency of patients who have the problem and produce a positive test (true positives) relative to all patients who produce positive tests (true positives and false positives).
Predictive value = $TP/(TP+FP)*100$

Efficiency is the percent of patients correctly classified as having the problem or not having the problem. This is what you really want to know!!!
Efficiency = $(TP+TN)/ALL TESTS * 100$

C. Factors effecting a test's efficiency: Lab tests tend to come back wrong a predictable percentage of the time due to operator, machine errors, gradual slips away from calibration, etc. When machines are calibrated, the settings get changed and calibration may not always be done exactly correctly. The “gold standards” machines are calibrated against tend to have variation and mistakes in them as well. Clinical laboratories frequently set their own “normal” values as several standard deviations away from the mass of results given by the test in question. Two labs may produce very different values even though they are both using the same machines because their gold standards and “norms” differ. When a false positive or false negative is suspected, ordering the test again is a traditional and appropriately conservative reaction.

When test results are being used for research, it is crucial that the investigator be very familiar with how the normal levels were set. What and how large was the population used to develop the norms, how were the tests done, what is the expected variability? Have “outlier” statistics been used to determine how likely a result of a particular value is to be abnormal? Does the cut off value for normal/abnormal make sense in light of the study's aims? What concurrent conditions, medications, etc. effect the test's values?

D. Interpreting the result in relation to the real world: If you take a test and get a positive result, what are the odds you have the disease? You can't make an interpretation in a vacuum. You need to consider the background of the patient being tested in order to guess whether the test's response is meaningful - or even if it is worth giving the test. For example, a young, healthy

person is unlikely to have a problem usually associated with old age and chronic ill health so a positive finding on a test is very likely to be a false positive.

Bayes Theorem (adapted from Gonick and Smith, 1993): How many of the people who test positive for a disease have it?. For example, if one of 1,000 people has a disease and there is a test for it which comes back positive 99% of the time if you really have (true positive) it but 2% of the time if you don't (false positive), what percent of the people who test positive have the disease?

Given the above, out of 1,000 people one should have the disease and should test positive (a 99% probability) but 20 of the remaining 999 uninfected people should test positive (a 2% false positive rate) and, thus, only 979 are left to test negative. Thus, less than 5% of the people who test positive actually have the disease. So, if you test positive, there is a 95% chance that you do not have the disease.

E. Testing differences between frequencies of occurrence and proportions - CHI Square

(X²) : Chi square techniques are used to evaluate changes in frequency of occurrence or in proportion of subjects responding. The Chi square test indicates how likely it is that the expected frequencies are different from those you observe. This test is used in many different ways. Three common uses are (a) to see if the actual distribution of proportions of responses is different than chance, (b) to see if the actual distribution is similar to a theoretical distribution, and (c) comparing distributions of proportions for two or more groups. For example, let's say that you used the protocol in pilot sample one to treat a mixture of patients with headaches. Your patients had either migraines, mixed tension and migraines, or tension headaches. At the start of the study, you would have no reason to think that the headaches would respond differently to the treatment. You would want to know if the actual proportion of responders differed from chance to see if the treatment was differentially effective depending on the type of headache. As there are three types of headache, you would expect the subjects showing excellent responses to be randomly divided between all three - or 1/3, 1/3, 1/3. So, your expected frequencies are all 033.3. Table 18 shows what a typical Chi Square matrix would look like if the actual frequencies of all headache subjects showing excellent responses were 70 migraine, 25 mixed, and 5 tension. When the Chi square statistic is calculated, the degrees of freedom equal the number of rows minus one times the number of columns minus one in the matrix. (DF = (r-1)X(c-1)). For this example, Chi square equals 35 with two degrees of freedom which is highly significant (p = at least 0.0001).

Thus, the proportions are significantly different than chance so the treatment is differentially effective.

Table 18 Chi Square matrix with equal expected frequencies.

	Migraine	Mixed	Tension
Actual frequency	70	25	5
Expected frequency	33	33	33

However, the protocol indicates a theory for how the treatment works. The theory involves changes in blood flow which should effect migraine headaches but not tension headaches. Thus, when the study is run, an equal proportion of the three groups would not be expected to respond excellently. Rather, only a few people with only tension headaches should respond excellently because they should be responding to a placebo. While up to a third of headache patients respond to placebos to some extent, only a very few respond excellently. If the people with mixed headaches have an average of half migraines and half tension headaches, they should respond about half as well as those with migraines. So, the expected frequency distribution would be something like 60, 30, 10 respectively. Table 19 shows what the chi square matrix would look like in this case. This time, Chi square is only 0.35 which is not significant. Thus, the actual proportions of responders were not different than the expected proportions so the theory is supported.

Table 19 Chi Square matrix with unequal expected proportions.

	Migraine	Mixed	Tension
Actual proportion	70	25	5
Expected proportion	60	30	10

It is frequently important to compare frequency of responses across groups. Let's continue using the sample protocol. Several different subsets of people with migraine headaches participated. One way to divide them up would be by other conditions related to their headaches. You might look at women who reported that headache onset was related to (a) their menstrual cycles, (b) changes in barometric pressure, or (c) no noticed factors. If you were examining those subjects who had an excellent response to the treatment but wanted to see if there were differences between the groups in how long it took to respond, you would use a 3X3 Chi square as shown in Table 20a. A more typical example is illustrated in Table 20b. In the first example, Chi square equals 93.3 with 4 DF which indicates that the proportions are different at $p =$ at least 0.0001. The second example also indicates a significant difference in proportions of frequencies as Chi square equals 249.8.

Table 20 3 X 3 factorial design for comparing frequencies across groups

20a: Time required to respond to PEMF vs. factors related to headache onset

Factors related to headache onset	# weeks of exposure required to respond to PEMF		
	after 1 week	after 2 weeks	after 3 weeks
menstrual	20	20	60
barometric pressure	40	30	30
unknown	80	15	5

20b: Type of music enjoyed vs. age range.

Age range of respondents	which type of music preferred		
	teen group	Barney	oldies
kids	40	50	10
teens	90	0.1	0.99
over the hills	20	0.0000000001	80

You can have many rows and columns if necessary. However, you can not have an expected frequency of less than five in any cell or the test will NOT produce meaningful results. Fisher's "exact" test does this for a 2 X 2 matrix.

F. Evaluation of relationships between direction of change - the Sign test: This test is used to see if two groups have changed in the same direction (e.g. up or down). For example, if patients with a disease are given a baseline measurement of severity and then divided into two groups, each receiving a different treatment, you could rate each patient at having improved (+) or not improved (-) (or better vs. worse). This rating would give you one plus or minus for each patient in each group. The sign test simply compares the number of signs in the groups and tells you whether the difference is likely to be significant.

Chapter 31

Outliers - data points that don't meet expectations

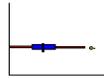
A: Concept: A major problem for researchers is what to do with those pesky numbers that don't fit your theory and fall far outside the range of most of your numbers. You can get rid of them if they are outside the limits of the test you are doing (and, thus, an obvious failure of the test) or if you can show in some other way that the number is simply incorrect. Otherwise you are stuck with it. Most people simply analyze and report the data both with and without these “outliers”. As illustrated in Figure 4 several pages back, you don't know if you happen to be looking at a very real response which is simply unique in your sample or a mistake. You can get some support for considering a data point an “outlier” by looking at how it sits relative to the other data points by taking into account both its location and variability of the other points.

B. Evaluating potential outliers in the distribution free setting: The interquartile range is one way to measure the spread of the data. The data points are put in sequential order and divided into four equal groups (equal number of points in each group). The three “quartiles” are the dividing lines between the quarters. The median of the high group (quartile 3) minus the median of the low group (quartile 1) gives you the interquartile range. Many statisticians (e.g. Gonick and Smith 1993) feel that a data point is an outlier if it is more than 1.5 times the size of the interquartile range above the third quartile or below the first. This is illustrated by a “box and whisker” plot (Figure 21 below) in which the ends of the box are the first and third quartiles with the group median drawn in between. Lines (the whiskers) extend 1.5 interquartile range lengths from both ends of the box. Any point beyond the whiskers would be considered an outlier.

Figure 21

A “box and whisker” plot used to identify outlier data points

The median is shown by the vertical line surrounded by the interquartile range. The whiskers are 1 ½ times the length of the interquartile range. The circle beyond them is an outlier point.



C. Evaluating outliers in the somewhat normally distributed setting: The zM test for outliers is used to see if an outlier is a real part of the rest of the sample or if a small sample is part of a larger, well known population (e.g. pulse rates of a sick group relative to a healthy matched norm).

Assumptions (read “requirements”) to use the test: normal distribution and random sampling.

Information needed:

n = number outliers

m = mean of outliers

M = mean of population

SD = standard deviation of POPULATION (not of the sample)

= absolute value

X = multiply

$$Z = \frac{(\text{square root } n) \times (M - m)}{\text{SD}}$$

Example: One outlier with the value of 100.

The rest of sample has a mean of 50 and a SD of 5. Z = 10.

Significance: For Z to be significant at the 0.05 level, it must be greater than 1.96 and at the 0.01 level greater than 2.58. Significance indicates a real outlier.

Chapter 32

Pattern analysis

A: Concept: Very often so many questions are asked that it is impossible to graph it out in a way which would permit the eye to detect patterns. This happens when two groups of patients are given a questionnaire containing dozens of questions with each question being answered on a multipoint scale (e.g., 1 to 10; good - fair - poor - horrid). A statistical technique called "pattern analysis" looks for patterns of question responses which would best differentiate between your groups (e.g. high numbers on questions 1, 7, & 32 with low numbers on 13 & 4 are related to group one while high numbers on 3, 13, and 19 with low numbers on 1, 6, and 17 are related to group two).

A technique called logistic regression / Probit analysis is used to determine which constellation of many factors best predict which of two groups a patient would fall into. Discriminate / Factor analysis is conceptually similar but is used when there are more than two groups which the patient could fall into.

B. Discriminate analysis and probit/logistic regression: These techniques are used when trying to see which combination of continuous and discontinuous variables best predicts a categorical outcome measure. Discriminate / Factor analysis is used when the grouping variable has more than two choices (a, b, and c) while probit analysis and logistic regression is used when there are only two choices (yes/no). For example, if you asked amputees about presence or absence of phantom pain (a yes/no response) you might want to know if questions about severity of stump pain, frequency of urination, war fought in, amount of prosthetic use, and amount of stress the respondent is under predict a yes or no response to having phantom pain. The analysis might tell you that very frequent stump pain and high prosthetic use tend to be related to report of phantom pain while low stress levels and having been in WWII are predictive of not reporting phantom pain. Frequency of urination might not be noted as helpful in predicting whether phantom pain was reported or not. Discriminate analysis would be used if you asked whether the phantom pain was burning, shocking, or cramping (3 categories) in order to see if any of your variables could predict which category a particular patient fell into.

When many variables are correlated with just two possible outcomes (e.g., getting better or not getting better - yes/no), you can not use the usual regression techniques because they are set up for continuous variables. Use discriminant analysis techniques instead. *An important drawback to these techniques is that their formulae require that each case have valid data for each variable or it is dropped from the analysis.* Thus, important cases can be left out of the analysis because one minor bit of data is missing. These techniques determine which combination of variables are best at predicting the outcome (account for most of the variability).

You have to determine which method the statistics program should use in selecting variables for the equation. I normally choose to minimize Wilk's lambda (group variability small relative to overall variability) and to maximize the Eigen value to emphasize differences between groups as relatively large compared to differences between subjects in the group. These are the usual settings for clinical studies. Do not permit the computer to choose the variables to use because it will give misleading information if two of the variables are very highly correlated. You need to tell it to skip one of them. The test begins by finding the one variable which is best at separating the groups. Then it adds the others one by one to see if any increase the ability to predict membership in a particular group. If two of the variables were highly correlated, both would have about the same ability to predict membership in a group. This would lead you to think you have a better chance of predicting group membership than you really do.

This is a very important test because you want to know if your variables can predict something such as whether a patient will live or die depending on which treatment you apply. For example, let's say you perform a discriminate analysis on two ways of handling a potentially a deadly disease. In one, you wait for symptoms to worsen and in the other, you begin treatment immediately. The analysis shows that the two group's discriminate scores differ at $p = 0.00001$ because the percent of patients correctly sorted (correctly predicted by the test) was 97.92%. That would make most investigators ecstatically happy. However, look at a possible matrix which produced this result:

	group 1	group 2
predicted	35	12
actual	35	13
% correct	100%	92.3%

The variables do predict outcome very accurately. However, in group two, one person out of 13 was missed. This person had a 50% chance of dying if the correct Rx was not started immediately. Thus, if you used this relatively accurate prediction, you would not treat 7.7 percent of your patients when they needed it and could wind up killing 3.9%. Thus, **this prediction is not accurate enough to use for this disease.**

C. Evaluating the contribution of various factors to predicting the outcome: Stepwise regression and logistic regression are similar to the above tests but are used with constantly varying measures to look for patterns of answers across the entire questionnaire which can differentiate between sub-groups of interest. Continuing the above example, you might use stepwise regression to see which variables predict high vs. low levels of phantom pain when the respondents rated their pain on a scale of zero to ten.

D. Determining which variables change together: Canonical and principal components analyses are used with discriminant analysis to reduce the number of variables which predict the classification by determining which variables change together (thus, you only have to look at the

one which gives the very best prediction). These techniques usually have a graph associated with them which shows how well the variables differentiate between the two outcome possibilities and highlight outliers which might be going the other way. It is also important to know which variables change together because they may be measuring the same thing.

E. Locating sub-groups of responses and respondents: Cluster analysis looks for patterns of responses which could lead to identification of distinct sub-groups of subjects. Most of the clusters are obvious nonsense due to chance high correlations of scores but some may point out previously unrecognized sub-groups in your population.

Chapter 33

Survival / life table analysis

Longitudinal studies have the severe weakness that people's lives frequently are not in concordance with progress of the study. They tend to die, leave town, etc. before the investigators get all the data they need. For example, if you want to know how long a hip implant will last, many of the patients will die of causes unrelated to the hip replacement before the implant wears out. Their data can not be analyzed by saying that the hip only lasted as long as they lived as this would be misleading. It would also be a real problem to discard the data from everyone who lives long enough for the implant to outlast them because the results would be biased toward implants which fail early. Exactly the same situation is faced by investigators trying to follow any group of patients in our highly mobile society. Long term follow-ups of behavioral interventions for chronic pain face exactly the same problem as the hip replacement studies not because the patients were elderly when they entered and die, but because they disappear.

The simplest way to make use of incomplete data of the sort described above is to plot out how many people's hip replacements hadn't failed, hadn't had their headaches return, etc. at specified periods of time (usually one year) after the intervention. This "survival" curve is essentially a plot of the survival rate against the time since the initial event. As it is unusual for all subjects to begin a longitudinal study at the same time, all of the start dates are set to an artificial "zero" so the longevity data can be compared. Table 21 provides sample data for a study in which chronic migraine headache patients were followed for five years after successful interventions. Thus, all began the follow-up period headache free.

Table 21 Follow-up of successfully treated chronic migraine headache patients
(100 subjects began the follow-up study)

Year after treatment for 100 subjects who were followed-up	number of subjects at risk that year	number of subjects whose headaches returned that year	number of subjects who could not be contacted that year
0 - 1	100	10	5
1 - 2	85	20	5
2 - 3	60	10	10
3 - 4	40	5	10
4 - 5	25	5	15

When a subject disappears at an unknown time during an interval, it is assumed that the person

disappeared in the middle of it. This assumption will be correct on the average if enough subjects participate in the trial. From the data in Table 21, 97.5 subjects were at risk of having their headaches return. This is because we give each of the drop-outs credit for a half year. The headaches returned for ten of the subjects so the risk of headaches returning during the first year is $10/97.5$ which is 0.10. The odds of surviving the year without headaches are $1 - 0.10$ which is 0.90.

For the second year, 82.5 people were at risk and 20 had their headaches return so the odds of remaining headache free are 0.76. The odds of getting through both years without headaches are the odds of the first year multiplied by the odds for the second year or 0.89×0.76 which equals 0.68.

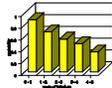
Continuing to the third year, 55 people were at risk and headaches returned for ten so the odds of not having headaches return during that year were 0.82. The odds of not having headaches return during the three year period were 0.56.

For the fourth year, 35 people were at risk and five had headaches return giving cumulative odds of 0.48. Similarly, for the fifth year the cumulative odds are 0.34.

The data can be plotted as the cumulative proportion of the subjects surviving (their headaches not returning) or as the probability of survival (headaches not returning). Figure 22 shows such a plot.

Figure 22 Survival curve for the data presented in Table 21

(yearly probabilities of headaches not returning during a five year follow-up)

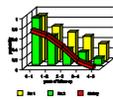


Frequently, the long-term results of two treatments need to be compared or the long-term results of a study need to be compared with the known natural history of the disorder. Powerful

statistical techniques for comparing two survival curves are available. However, the simplest route is frequently to straighten the curves using the transformation techniques discussed earlier and applying a standard formula for comparing them. Common techniques include the Mantel-Haenzel Chi-square and the Gehan-Wilcoxon test. If you want to find out which variables influence the regression line, use the Cox proportional hazards regression technique. Good statistics packages have either these or equivalent techniques. A typical situation in which three survival curves are available for similar subjects is presented in Figure 23.

Figure 23 Survival curves for two treatments and natural history

(yearly probabilities of headaches not returning during a five year follow-up)



Chapter 34

Establishing Efficacy: Using effect sizes to determine how powerful a treatment is likely to be

This section contains two major parts:

- (1) reviews vs. meta-analysis and effect size and
- (2) formal evaluation methodology for establishing effectiveness

The key questions addressed here are: How do we figure out how effective a treatment (such as EMG biofeedback to prevent tension headaches) actually is? How do we objectively compare the efficacy of several different treatments (e.g. Temp BFB vs. Propranolol to prevent migraines)?

1. Old style literature reviews vs. meta-analysis based effect size determinations

We all used to review the literature and take a best guess.

Now, we perform statistics as part of doing a “meta-analysis” to produce results phrased using such terms as “effect size”.

When we used to review the literature on some topic (such as my old review of biofeedback for low back pain), we would find every article we could on the topic and essentially make a big table listing how many subjects there were, what the treatment was, how many folk got how much better, follow-up duration, etc. We tried to synthesize the information into a “best guess” about how well the treatment worked by informally factoring in study quality (clear diagnosis, controls, well defined Rx, meaningful / objective outcome measures, etc.) and trying to match designs as well as we could.

Our conclusions were always rough and open to considerable bias.

First off – what is a **Meta-Analysis**?

The limitations of reviews are somewhat compensated for by including “meta-analytic” techniques.

The idea seems great: Use statistics to combine a bunch of small, relatively weak studies into one big, relatively strong study with sufficient patients to be more certain of how efficacious the outcome is.

We begin by finding all the studies on the topic (e.g. BFB for back pain).

Then we do some quality control by eliminating all the studies which are very weak (e.g. no clear Dx, can't tell what the Rx really was, no decent outcome measures, etc.).

Next, we combine the data from studies with similar designs which use similar treatments and similar objective outcome measures.

This eliminates most of the studies but leaves us with relatively strong, comparable

studies.

After doing magic statistics (which most of us need help from a professional statistician to perform correctly), we come up with numbers showing how effective the treatment probably really is and how certain we are of that judgment.

The most relevant to us is “effect size”.

Effect Size - For most of us, the term “effect size” brings to mind something about how effective treatments are judged to be after a meta-analysis.

Just how Effect Size is calculated, what it is to do with a meta-analysis, and how it relates to the more familiar “significance values” tends to be pretty cloudy.

The key concept to remember is that **Effect Size is how effective the treatment actually is**. This means we can:

- a. objectively determine the effectiveness of a treatment and use it to predict how well patients are likely to do and
- b. directly compare the effectiveness of two treatments

The following is Modified from “Research Rundowns” web site Sept 13

“What is Effect Size?”

The simple definition of effect size is the magnitude, or size, of an effect. Statistical significance (*e.g.*, $p < .05$) tells us there was a difference between two groups or more based on some treatment or sorting variable. What this fails to tell us is the magnitude of the difference. In other words, *how much more effective* was (treatment one than treatment two)?

To answer this question, we standardize the difference and compare it to 0 – no effect.”

“One type of effect size, the standardized mean effect, expresses the mean difference between two groups in standard deviation units.”

Typically, you’ll see this reported as Cohen’s *d*, or simply referred to as “*d*.”

“How can effect sizes be interpreted?”

One feature of an effect size is that it can be directly converted into statements about the overlap between the two samples in terms of a comparison of percentiles.

An effect size is exactly equivalent to a 'Z-score' of a standard Normal distribution.

For example, an effect size of 0.8 means that the score of the average person in the experimental group is 0.8 standard deviations above the average person in the control group, and hence exceeds the scores of 79% of the control group.”

(From: It's the Effect Size, Stupid What effect size is and why it is important. Robert Coe School of Education, University of Durham, email r.j.coe@dur.ac.uk
Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, 12-14 September 2002)

For us, the effect size is stating the objective the difference between the means & standard deviations for two treatments or a treatment vs. placebo. It tells us how much difference in outcome we can expect between a person in the control group and the treatment group.

The values calculated for effect size are generally in the range of 0 to 3.0

The meaning of effect size varies by context, but the standard interpretation offered by Cohen

(1988) is:

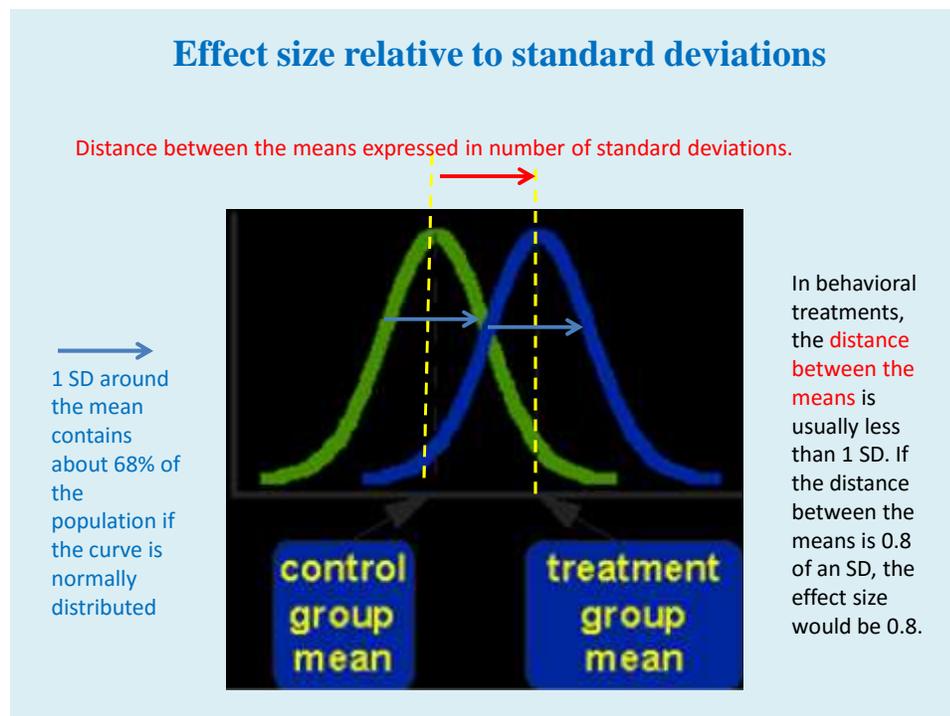
.8 = large (8/10 of a standard deviation unit)

.5 = moderate (1/2 of a standard deviation)

.2 = small (1/5 of a standard deviation)”

Biofeedback based interventions usually show effect sizes in the range of 0.4 to 0.8. If biofeedback treatments are effective, why are the effect sizes so small? Shouldn't they be at least 2 instead of a pitiful 0.4 to 0.8? Not really. An Effect size of 2 means that an average of 98% of the placebo controls can be differentiated from the real treatment group. Biofeedback based interventions help about 60 to 80 percent of patients to show improvements of about 80% in various measures of headache activity such as frequency, duration, intensity, debilitation, and drug use. Good HA placebos get the same results for about 30 percent of matched patients. (I once got a placebo response among 50% of patients showing at least minor improvements for at least six months). Thus, our effect sizes are realistic.

The following slide (Sherman 2013) explains how effect size is related to the distance between the means of two groups in number of standard deviations.



Effect Size **Percentage of control group who would be below the average person in the experimental group**

0.0	50%
0.1	54%
0.2	58%
0.3	62%
0.4	66%
0.5	69%
0.6	73%
0.7	76%
0.8	79%
0.9	82%
1.0	84%
1.2	88%
1.4	92%
1.6	95%
1.8	96%
2.0	98%
2.5	99%
3.0	99.9%

2. Method of rating treatment efficacy established by AAPB and ISNR:

a. Efficacy vs. Clinical Effectiveness:

Efficacy is determined by evaluating formal studies done on each disorder. When a study is done, the treatment is very carefully standardized, the people doing the interventions should have great expertise in the treatment and the disorder, and patients are very carefully selected. In the real clinical environment, the patients may have many problems in addition to the one they are being treated for (which would affect the chances of the treatment doing well), may be given many overlapping treatments at once (so you can't tell how much help any one treatment was), and the therapist may not be as experienced as the people running the research study. Thus, a treatment's efficacy may be greater or lesser than its effectiveness in the real clinical world.

b. Establishment of Rating Criteria:

The Association for Applied Psychophysiology has developed the following criteria for setting the level of evidence for efficacy (Moss and Gunkelman 2002, LaVaque et al 2002): It is very similar to the rating schemes developed by other organizations such as the American Psychological Association. Please note that the efficacy ratings made based on these criteria are from formal studies. Please see these citations for an explanation of how the ratings were arrived at and a discussion of the weaknesses of double blind studies for several of the techniques evaluated.

LaVaque, T., Hammond, D., Trudeau, D., Monastra, V., Perry, J., Lehrer, P., Matheson, D., & Sherman, R. (2002). Template for developing guidelines for the evaluation of the clinical efficacy of psychophysiological evaluations. *Applied Psychophysiology and Biofeedback*, 27(4), 273–281. Co-published in *Journal of Neurotherapy*, 6, 11–23.

Moss, D. & Gunkelman, J. (2002). Task force report on methodology and empirically supported treatments: Introduction. *Applied Psychophysiology and Biofeedback*, 27, 261–262.

c. Criteria:

(Quoted directly from Yucha and Montgomery's Evidence-Based Practice in Biofeedback and Neurofeedback, 2008)

“Biofeedback therapy has matured over the last 30 years, and today there are myriad disorders for which biofeedback therapy has been used. Large research grants have funded prospective studies on biofeedback therapy for a variety of disorders, such as headache (migraine, mixed, and tension), essential hypertension, and urinary incontinence. These studies consistently report positive results.

On the other hand, several reports of unsuccessful biofeedback training have appeared in the research literature since the inception of biofeedback training three decades ago. Many of the unsuccessful studies conducted in the early development of the field reflect failure to thoroughly train patients. For example, some unsuccessful studies provided only minimal training with the

biofeedback instrumentation (often one to four sessions of short duration), provided little coaching, involved no home practice, and failed to train to clinical criteria.

In 2001, a Task Force of the Association for Applied Psychophysiology and Biofeedback and the Society for Neuronal Regulation developed guidelines for the evaluation of the clinical efficacy of psychophysiological interventions (Moss & Gunkelman, 2002). The board of directors of both organizations subsequently approved these guidelines.

These Criteria for Levels of Evidence of Efficacy, described below, were used to assign efficacy levels for the vast number of conditions for which biofeedback has been used.

Level 1: Not Empirically Supported

Supported only by anecdotal reports and/or case studies in nonpeer-reviewed venues. Not empirically supported.

Level 2: Possibly Efficacious

At least one study of sufficient statistical power with well-identified outcome measures but lacking randomized assignment to a control condition internal to the study.

Level 3: Probably Efficacious

Multiple observational studies, clinical studies, wait-list controlled studies, and within-subject and intrasubject replication studies that demonstrate efficacy.

Level 4: Efficacious

- a. In a comparison with a no-treatment control group, alternative treatment group, or sham (placebo) control utilizing randomized assignment, the investigational treatment is shown to be statistically significantly superior to the control condition, or the investigational treatment is equivalent to a treatment of established efficacy in a study with sufficient power to detect moderate differences, and
- b. The studies have been conducted with a population treated for a specific problem, for whom inclusion criteria are delineated in a reliable, operationally defined manner, and
- c. The study used valid and clearly specified outcome measures related to the problem being treated, and
- d. The data are subjected to appropriate data analysis, and
- e. The diagnostic and treatment variables and procedures are clearly defined in a manner that permits replication of the study by independent researchers, and
- f. The superiority or equivalence of the investigational treatment has been shown in at least two independent research settings.

Level 5: Efficacious and Specific

Evidence for Level 5 efficacy meets all of the criteria for Level 4. In addition, the investigational treatment has been shown to be statistically superior to credible sham therapy, pill, or alternative bona fide treatment in at least two independent research settings.

In this particular update, we asked a professional librarian (Eva Stowers, University of Nevada, Las Vegas) to provide a comprehensive literature search of biofeedback and neurofeedback articles.

Criteria used included being published in a peer-reviewed journal between 2003 – 2007. When

there were numerous higher level research studies available, case studies were not added to this version. Abstracts and articles in languages other than English were not included. This monograph is not meant to be an inclusive review of all literature published on every possible disorder, but rather is meant to provide rationale for efficacy ratings of biofeedback.

References

Moss, D., & Gunkelman, J. (2002). Task force report on methodology and empirically supported treatments: Introduction and summary. *Biofeedback*, 30(2), 19-20.

Moss, D., & Gunkelman, J. (2002). Task force report on methodology and empirically supported treatments: Introduction and summary. *Applied Psychophysiology and Biofeedback*, 27(4), 261-262.”

Chapter 35

Problems establishing cause and Effect

How do you figure out what causes a problem / disorder?

I. The five key concepts are:

1. Double blind, placebo controlled therapeutic trials and most repeated measures designs don't determine cause.
2. Longitudinal designs are used to determine cause. This design permits determination that changes in a proposed causative variable come before and match changes in the problem.
3. Many different cascades of antecedents can lead to one proximal cause leading to one symptom.
4. The actual sequence of physical events from initial stimulation all the way through brain reactions to a consciously reported symptom can be recorded.
5. Correlation – even a very high one – does not provide evidence of cause. Biomarkers are correlates and / or markers rather than causes.

A. Double blind, placebo controlled therapeutic trials are not meant to establish what is causing the problem being investigated. All that a well-designed study of this type can tell us is that the problem really did change from before to after some intervention while the problem did not change due to extraneous factors such as time and unanticipated intervening variables as the placebo group doesn't change as much as the intervention group. The importance of using a very believable placebo, having truly blinded / neutral evaluators, having all subjects rate whether they were in the placebo or experimental group, and other crucial design considerations are not relevant to this discussion.

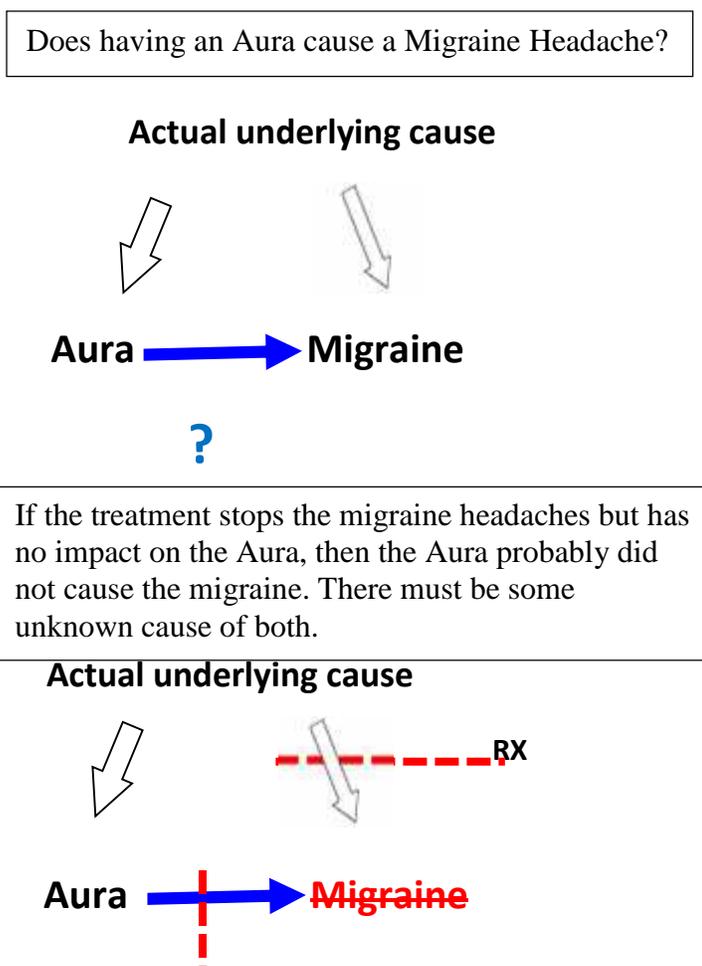
B. Longitudinal study designs – either with or without a placebo - are the most powerful way to demonstrate that consistent, repeated changes in an underlying factor causes parallel changes in the problem being investigated. Changes in the proposed cause must come **before** all changes in the problem and match changes in intensity.

1. Longitudinal studies have to be carefully designed so some intermediate variable or co-varying factor is not mistaken for the cause. For example people with chronic migraine headaches may or may not get auras before or during a migraine episode. Many people get auras identical to those associated with migraines for decades but never have migraine headaches. Several studies, including three by our team (Sherman 2012) showed that effective interventions for migraine frequently do not result in any change in auras. Thus, some as yet unknown “cause” results in both auras and migraines – but one does not cause the other.

Longitudinal studies must be designed so any potential co-occurring variables are recorded using logs. Examples are detailed in the two papers presented near the end of this document.

A proposed causative variable must reliably change in both presence and intensity **before** the problem changes (a) by either starting or stopping before the problem does or (b) changes in intensity. In other words, it must **predict** any changes in the problem.

The following figure illustrates the concept that the existence of a common cause can be demonstrated by cutting one limb with no change in the other. In the following example, this means that auras do not cause migraines.



2. Many longitudinal designs cannot differentiate between a correlate and a cause because they cannot show that one variable consistently changes before the other. Is the following longitudinal study of burning phantom limb pain with repeated measures able to

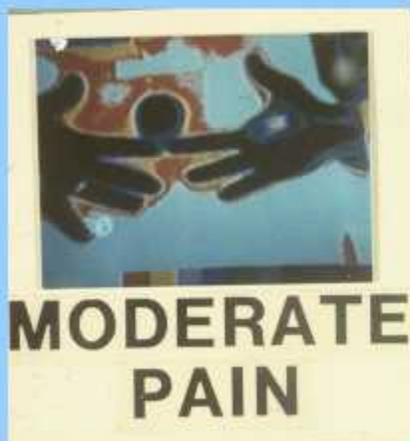
demonstrate a cause? No because readings were taken as a cross-section in time so there is no way to know which changed first.

Finger amputee recorded four times with different intensities of burning phantom pain - slide 1 of 2.



The photos show darker colors as cool with white as cold and black as hot.

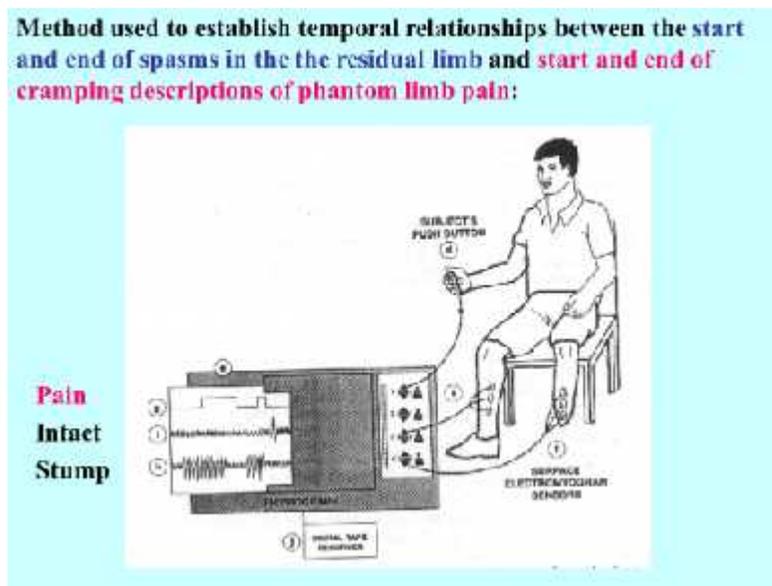
Finger amputee recorded four times with different intensities of burning phantom pain - slide 2 of 2.



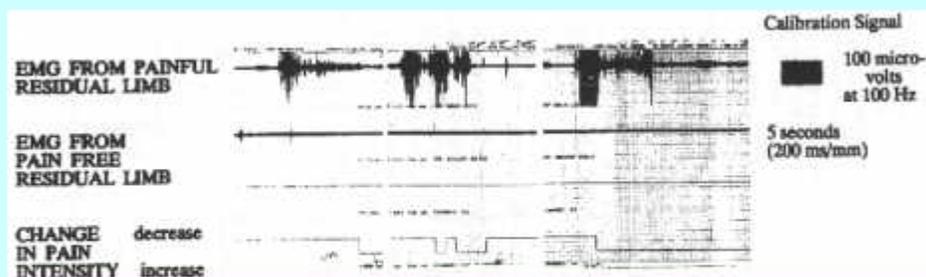
3. Examples of Studies from Sherman (2012) which demonstrate cause as they can predict changes: (Sherman R: Pain Assessment and Intervention from a Psychophysiological Perspective – Second Edition. Association for Applied Psychophysiology, Wheat Ridge Colorado, 2012. Details of each example are in the book. **Psychophysiology students can request an electronic copy from Dr. Sherman at no charge.**)

(a) Change in tension in the residual limb leading to cramping
Phantom pain.

(i) Longitudinal recordings over hours.



Sample from a chart recording showing surface EMG from the residual limbs of a bilateral amputee experiencing cramping phantom limb pain only in one of the residual limbs.



If the increased tension was a reflex reaction to the pain it would come after the pain began and there would be a tiny reaction in the pain-free limb as well.

Reaction time was calculated and was not the reason the cramp came before the button press indicating start of the pain.

(ii) Longitudinal studies in the environment over days showing consistent changes in muscle tension before changes in pain.

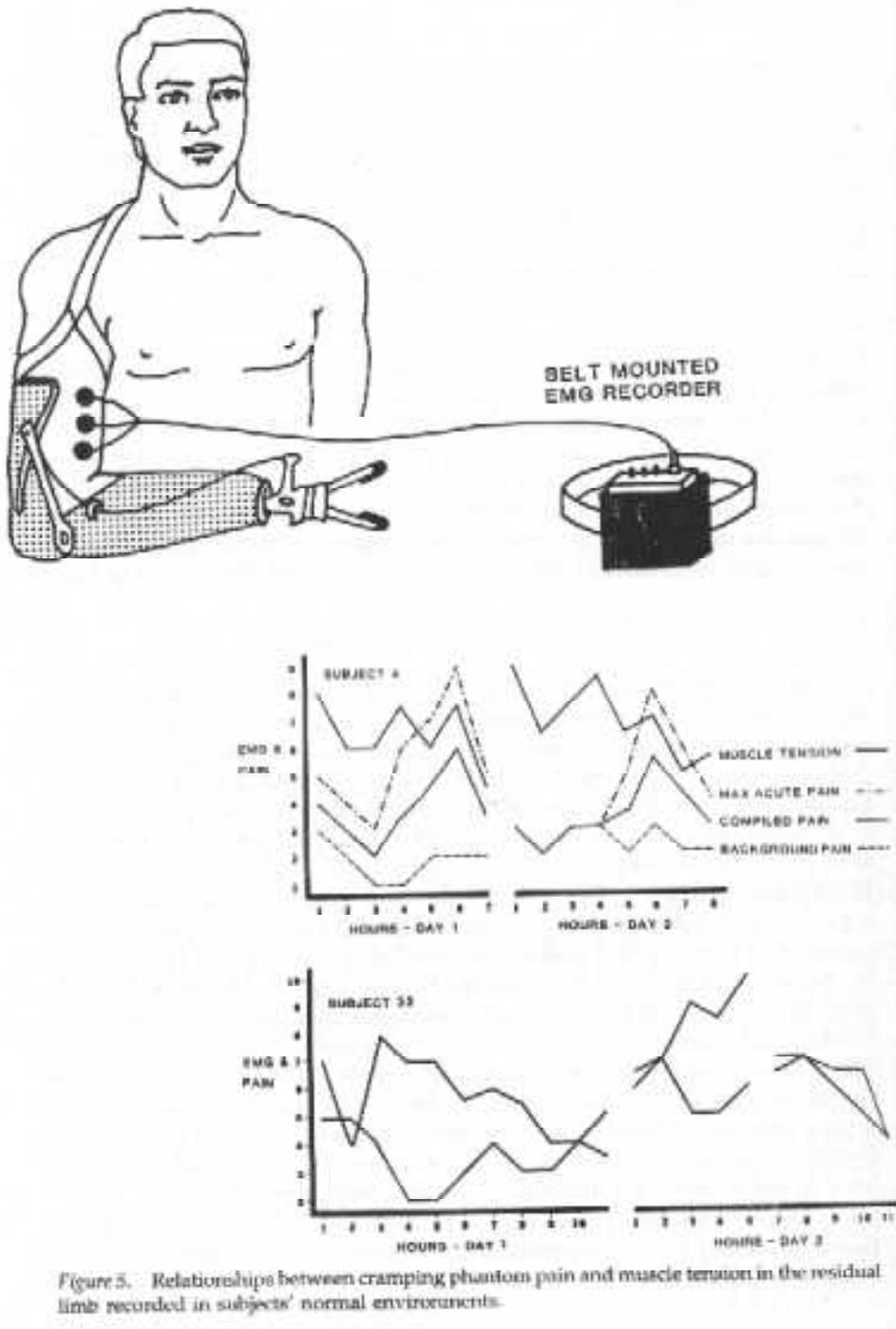


Figure 5. Relationships between cramping phantom pain and muscle tension in the residual limb recorded in subjects' normal environments.

(b) Predictive changes in neck / upper back tension resulting in tension headache.

The figures below show the relationships between shoulder muscle tension and tension headache intensity recorded in a subject's normal environment before and after a combination of sEMG biofeedback and progressive muscle relaxation training.

Activity during the two four-hour periods was similar. Height of vertical deflection is indicative of amount of muscle tension.

After training, sEMG is not only lower but quiet periods of exceedingly low tension appear in the recording.

Before training (pain ranged from 2 to 3 on a zero to ten scale)

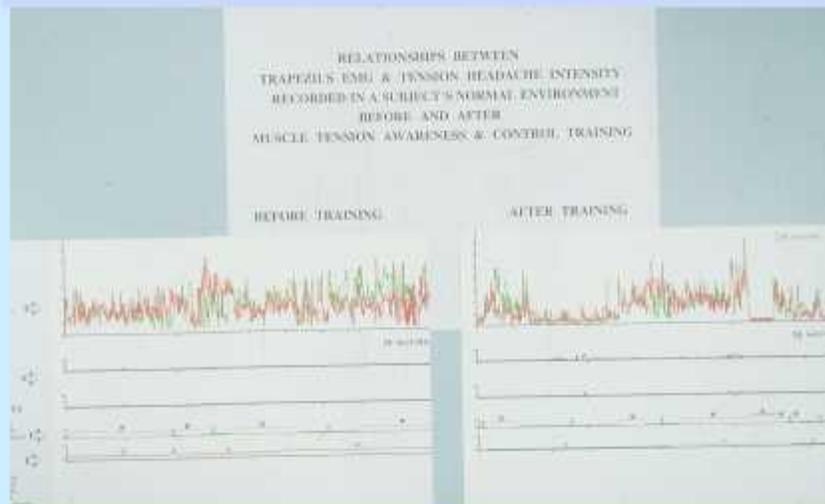


After training (pain ranged from 0 to 1 on a zero to ten scale)



Patients who did not learn to lower their muscle tension did not show decreased headache / pain.

sEMG can be used to track changes in pain - muscle tension relationships over time.



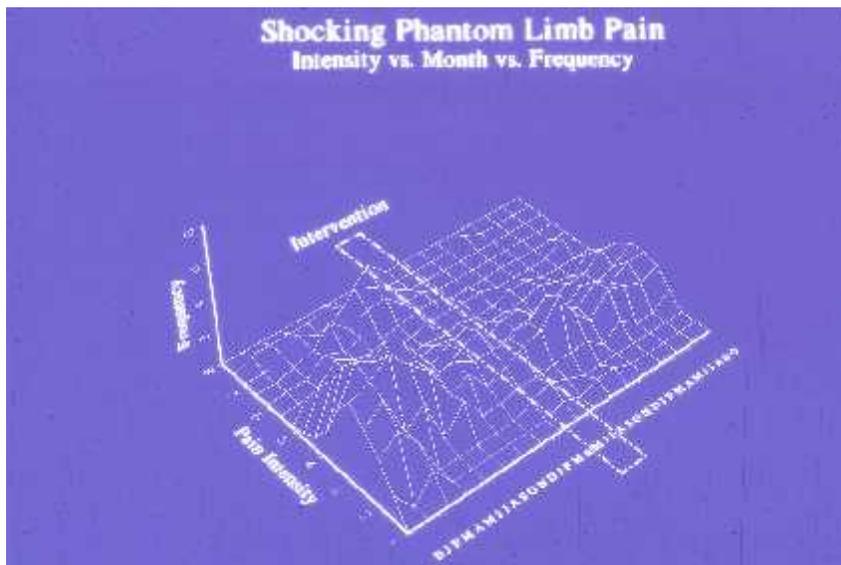
(c) Change in blood flow in the residual limb leading to burning phantom pain - change in firing of neuroma recorded.

(i) Experimental manipulation of blood flow by motion and by blocks which only alter blood flow. Blood flow and tension were measured in the

following subject. Blood flow in the subject's limb was restricted when the limb was raised by the researcher. There was no change in muscle tension but blood flow decreased and then pain increased.

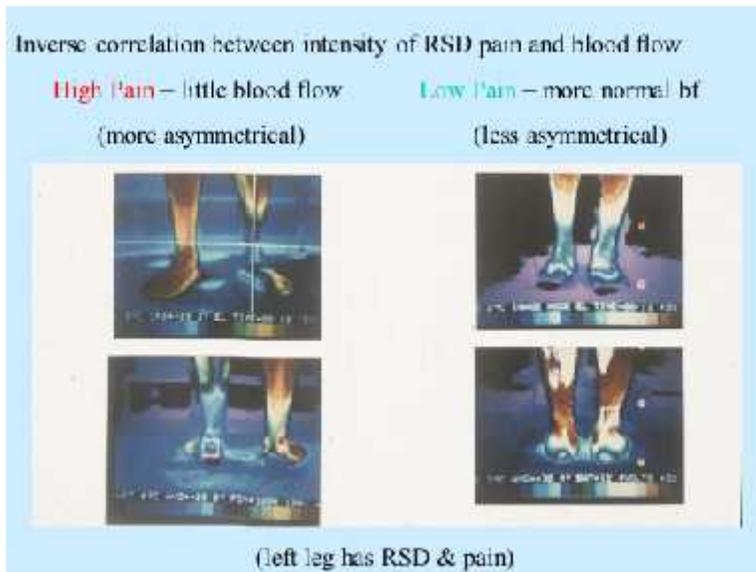


(d) Change in temperature leading to change in shooting phantom pain
Over 3 years - change in firing of neuroma recorded.



(e) Change in location of symptoms of RSD linked exactly with preceding changes in blood flow at the same location.

(i) In lab repeated measures.



(ii) Three year longitudinal recording of patient with RSD
Following changes in blood flow and pain through environmental and interventional changes.

(f) Stress – tension – pain relationships

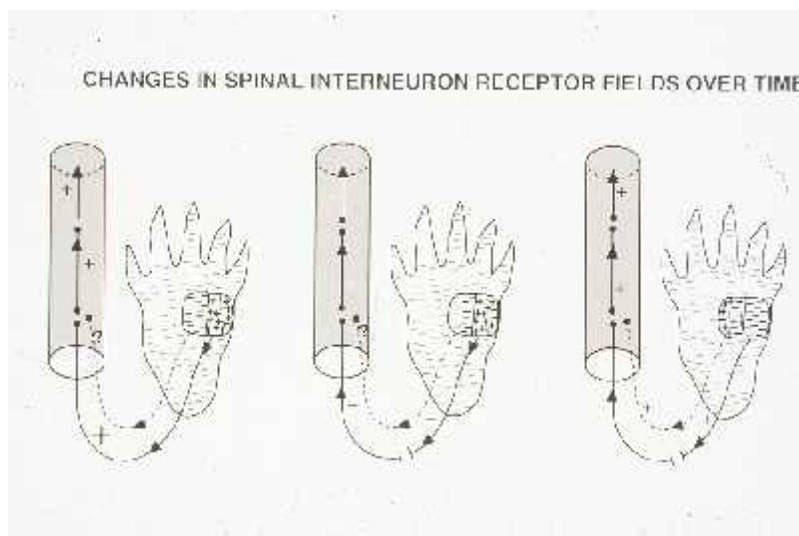


C. Many different cascades of antecedents can lead to one proximal cause

leading to one symptom.

We know that muscles kept too tense (even just 5%) for too long (even just 10 minutes) begin to hurt and that the pain may not resolve for hours. The underlying mechanism for this cellular metabolism induced pain has been established and followed all the way to the brain. Several scenarios can produce the same symptom: (a) The muscles might be kept tight because the person is not aware of the tension or (b) because they are tensing as a stress reaction which leads to increased sympathetic outflow which, in turn, causes spindle fibers to tense, which, in turn, causes the major muscles to tense.

D. The actual sequence of physical events from initial stimulation all the way through brain reactions to a consciously reported symptom can be recorded.



E. Correlates do not prove that one change in one variable causes change in another. Biomarkers are correlates or indicators.

1. A reliable biomarker for changes in spelling ability in a middle-class elementary school in the US is foot size. However, changes in foot size probably do not cause nor predict changes in spelling ability. Even a perfect correlation of plus or minus one does not indicate cause.

2. Biomarkers: Potential Uses and Limitations

[Richard Mayeux](#) NeuroRx. 2004 Apr; 1(2): 182–188.

Biomarkers provide a dynamic and powerful approach to understanding the spectrum of neurological disease with applications in observational and analytic epidemiology, randomized clinical trials, screening and diagnosis and prognosis. Defined as alterations in the constituents of tissues or body fluids, these markers offer the means for homogeneous classification of a disease and risk factors, and they can extend our base information about the underlying pathogenesis of disease. Biomarkers can also reflect the entire spectrum of disease from the earliest

manifestations to the terminal stages. This brief review describes the major uses of biomarkers in clinical investigation. Careful assessment of the validity of biomarkers is required with respect to the stage of disease. Causes of variability in the measurement of biomarkers range from the individual to the laboratory. Issues that affect the analysis of biomarkers are discussed along with recommendations on how to deal with bias and confounding.

Biological markers (biomarkers) have been defined by Hulka and colleagues¹ as “cellular, biochemical or molecular alterations that are measurable in biological media such as human tissues, cells, or fluids.” More recently, the definition has been broadened to include biological characteristics that can be objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention.² In practice, biomarkers include tools and technologies that can aid in understanding the prediction, cause, diagnosis, progression, regression, or outcome of treatment of disease. For the nervous system there is a wide range of techniques used to gain information about the brain in both the healthy and diseased state. These may involve measurements directly on biological media (e.g., blood or cerebrospinal fluid) or measurements such as brain imaging which do not involve direct sampling of biological media but measure changes in the composition or function of the nervous system.

3. From Wikipedia 19 May 15: In [medicine](#), a biomarker can be a traceable substance that is introduced into an organism as a means to examine organ function or other aspects of health. For example, [rubidium chloride](#) is used as a radioactive isotope to evaluate perfusion of heart muscle. It can also be a substance whose detection indicates a particular disease state, for example, the presence of an [antibody](#) may indicate an [infection](#). More specifically, a biomarker indicates a change in expression or state of a protein that correlates with the risk or progression of a disease, or with the susceptibility of the disease to a given treatment.

[Biochemical](#) biomarkers are often used in [clinical trials](#), where they are derived from bodily fluids that are easily available to the early phase researchers. A useful way of finding genetic causes of diseases such as [schizophrenia](#) has been the use of a special kind of biomarker called an [endophenotype](#).

Other biomarkers can be based on measures of the electrical activity of the brain (using [Electroencephalography](#) (so-called [Quantitative electroencephalography \(qEEG\)](#)) or [Magnetoencephalography](#)), or volumetric measures of certain brain regions (using [Magnetic resonance imaging](#)) or [saliva testing](#) of natural metabolites, such as saliva nitrite, a surrogate marker for [nitric oxide](#).

One example of a commonly used biomarker in medicine is [prostate-specific antigen](#) (PSA). This marker can be measured as a proxy of prostate size with rapid changes potentially indicating cancer.

4. Antidepressant response trajectories and quantitative electroencephalography (QEEG) biomarkers in major depressive disorder

[Aimee M. Hunter](#),^{a,*} [Bengt O. Muthén](#),^b [Ian A. Cook](#),^a and [Andrew F. Leuchter](#)^a

J Psychiatr Res. 2010 Jan; 44(2): 90–98.

Individuals with Major Depressive Disorder (MDD) vary regarding the rate, magnitude and stability of symptom changes during antidepressant treatment. Growth mixture modeling (GMM)

can be used to identify patterns of change in symptom severity over time. **Quantitative electroencephalographic (QEEG)** cordance within the first week of treatment has been associated with endpoint clinical outcomes but has not been examined in relation to patterns of symptom change. Ninety-four adults with MDD were randomized to eight weeks of double-blinded treatment with fluoxetine 20 mg or venlafaxine 150 mg ($n = 49$) or placebo ($n = 45$). An exploratory random effect GMM was applied to Hamilton Depression Rating Scale (Ham-D₁₇) scores over 11 timepoints. Linear mixed models examined 48-h, and 1-week changes in QEEG midline-and-right-frontal (MRF) cordance for subjects in the GMM trajectory classes. Among medication subjects an estimated 62% of subjects were classified as responders, 21% as non-responders, and 17% as symptomatically volatile—i.e., showing a course of alternating improvement and worsening. MRF cordance showed a significant class-by-time interaction ($F_{(2,41)} = 6.82, p = .003$); as hypothesized, the responders showed a significantly greater 1-week decrease in cordance as compared to non-responders (mean difference = $-.76$, Std. Error = $.34$, $df = 73, p = .03$) but not volatile subjects. Subjects with a volatile course of symptom change may merit special clinical consideration and, from a research perspective, may confound the interpretation of typical binary endpoint outcomes. Statistical methods such as GMM are needed to identify clinically relevant symptom response trajectories.

II. Two articles which detail factors which need to be controlled for when establishing cause of a problem

The following articles discuss how longitudinal studies need to be designed to defend against many complications. They are well worth reading.

A. Cause and Effect

by [Martyn Shuttleworth](https://explorable.com/cause-and-effect) <https://explorable.com/cause-and-effect> 16 May 15

Cause and effect is one of the most commonly misunderstood concepts in science and is often misused by lawyers, the media, politicians and even scientists themselves, in an attempt to add legitimacy to research.

The basic principle of causality is determining whether the results and trends seen in an [experiment](#) are actually caused by the manipulation or whether some other factor may underlie the process.

Unfortunately, the media and politicians often jump upon scientific results and proclaim that it conveniently fits their beliefs and policies. Some scientists, fixated upon 'proving' that their view of the world is correct, leak their results to the press before allowing the [peer review process](#) to check and [validate](#) their work.

Some examples of this are rife in alternative therapy, when a group of scientists announces that they have found the next healthy superfood or that a certain treatment cured swine flu. Many of these claims deviate from the scientific process and pay little heed to cause and effect, diluting

the claims of genuine researchers in the field.

What is Cause and Effect? - The Temporal Issue

The key principle of establishing [cause and effect](#) is proving that the effects seen in the experiment happened after the cause.

This seems to be an extremely obvious statement, but that is not always the case. Natural phenomena are complicated and intertwined, often overlapping and making it difficult to establish a natural order. Think about it this way: in an experiment to study the effects of depression upon alcohol consumption, researchers find that people who suffer from higher levels of depression drink more, and announce that this [correlation](#) shows that depression drives people to drink.

However, is this necessarily the case? Depression could be the cause that makes people drink more but it is equally possible that heavy consumption of alcohol, a depressant, makes people more depressed. This type of classic 'chicken and egg' argument makes establishing causality one of the most difficult aspects of [scientific research](#). It is also one of the most important factors, because it can misdirect scientists. It also leaves the research open to manipulation by interest groups, who will take the results and proclaim them as a truth.

With the above example, an alcoholic drink manufacturer could use the second interpretation to claim that alcohol is not a factor in depression and that the responsibility is upon society to ensure that people do not become depressed. An anti-alcohol group, on the other hand, could claim that alcohol is harmful and use the results to lobby for harsher drinking laws. The same research leads to two different interpretations and, the answer given to the media can depend upon who funds the work.

Unfortunately, most of the general public are not scientists and cannot be expected to filter every single news item that they read for quality or delve into which group funded research. Even respected and trusted newspapers, journals and internet resources can fall into the causality trap, so marketing groups can influence perceptions.

What is Cause and Effect? - The Danger of Alternative Explanations

The other problem with causality is that a researcher cannot always guarantee that their particular [manipulation](#) of a variable was the sole reason for the perceived trends and correlation.

In a complex experiment, it is often difficult to isolate and neutralize the influence of [confounding variables](#). This makes it exceptionally difficult for the researcher to state that their treatment is the sole cause, so any research program must contain measures to establish the cause and effect relationship.

In the physical sciences, such as physics and chemistry, it is fairly easy to establish causality, because a good experimental design can neutralize any potentially confounding variables. Sociology, at the other extreme, is exceptionally prone to causality issues, because individual humans and social groups vary so wildly and are subjected to a wide range of external pressures and influences.

For results to have any meaning, a researcher must make causality the first priority, simply because it can have such a devastating effect upon validity. Most [experiments](#) with some [validity](#) issues can be salvaged, and produce some usable data. An experiment with no established cause and effect, on the other hand, will be practically useless and a waste of resources.

How to Establish Cause and Effect

The first thing to remember with causality, especially in the non-physical sciences, is that it is impossible to establish complete causality.

However, the magical figure of 100% proof of causality is what every researcher must strive for, to ensure that a group of their peers will accept the results. The only way to do this is through a strong and well-considered experimental design, often containing pilot studies to establish cause and effect before plowing on with a complex and expensive study.

The temporal factor is usually the easiest aspect to neutralize, simply because most experiments involve administering a treatment and then observing the effects, giving a [linear](#) temporal relationship. In experiments that use historical data, as with the drinking/depression example, this can be a little more complex. Most researchers performing such a program will supplement it with a series of individual case studies, and interviewing a selection of the [participants](#), in depth, will allow the researchers to find the order of events.

For example, interviewing a sample of the depressed heavy drinkers will establish whether they felt that they were depressed before they started drinking or if the depression came later. The process of establishing cause and effect is a matter of ensuring that the potential influence of 'missing variables' is minimized.

One notable example, by the researchers Balnaves and Caputi, looked at the academic performance of university students and attempted to find a [correlation](#) with age. Indeed, they found that older, more mature students performed [significantly](#) better. However, as they pointed out, you cannot simply say that age causes the effect of making people into better students. Such a simplistic assumption is called a spurious relationship, the process of 'leaping to conclusions.'

In fact, there is a whole host of reasons why a mature student performs better: they have more life experience and confidence, and many feel that it is their last chance to succeed; my graduation year included a 75-year-old man, and nobody studied harder! Mature students may well have made a great financial sacrifice, so they are a little more determined to succeed. Establishing cause and effect is extremely difficult in this case, so the researchers interpreted the results very carefully.

Another example is the idea that because people who eat a lot of extra virgin olive oil live for longer, olive oil makes people live longer. While there is some truth behind this, you have to remember that most regular olive oil eaters also eat a Mediterranean diet, have active lifestyles, and generally less stress. These also have a strong influence, so any such research program should include studies into the effect of these - this is why a research program is not always a single experiment but often a series of experiments.

History Threats and Their Influence Upon Cause and Effect

One of the biggest threats to [internal validity](#) through incorrect application of cause and effect is the 'history' threat.

This is where another event actually caused the effect noticed, rather than your treatment or manipulation. Most researchers perform a pre-test upon a group, administer the treatment and then measure the post-test results ([pretest-posttest-design](#)). If the results are better, it is easy to assume that the treatment caused the result, but this is not necessarily the case.

For example, take the case of an educational researcher wishing to measure the effect of a new teaching method upon the mathematical aptitude of students. They pre-test, teach the new program for a few months and then posttest. Results improve, and they proclaim that their program works.

However, the research was ruined by a historical threat: during the course of the research, a major television network released a new educational series called 'Maths made Easy,' which most of the students watched. This influenced the results and compromised the [validity](#) of the experiment.

Fortunately, the solution to this problem is easy: if the researcher uses a two group [pretest-posttest design](#) with a [control group](#), the control group will be equally influenced by the historical event, so the researcher can still establish a good baseline. There are a number of other 'single group' threats, but establishing a good control driven study largely eliminates these threats to causality.

Social Threats and Their Influence Upon Cause and Effect

Social threats are a big problem for social researchers simply because they are one of the most difficult of the threats to minimize. These types of threats arise from issues within the participant groups or the researchers themselves. In an educational setting, with two groups of children, one treated and one not, there are a number of potential issues.

- **Diffusion or Imitation of Treatment:**
With this threat, information travels between groups and smoothes out any differences in the results. In a school, for example, students mix outside classes and may swap information or coach the [control group](#) about some of the great new study techniques that they have learned. It is practically impossible and extremely unfair to expect students not to mix, so this particular threat is always an issue.
- **Compensatory Rivalry:**
Quite simply, this is where the control group becomes extremely jealous of the treatment group. They might think that the research is unfair, because their fellow students are earning better grades. As a result, they try much harder to show that they are equally as clever, reducing the difference between the two groups.
- **Demoralization and Resentment:**
This jealousy may have the opposite effect and manifest as a built up resentment that the other group is receiving favorable treatment. The [control group](#), quite simply, gives up and does not bother trying and their grades plummet. This makes the educational program appear to be much more successful than it really is.

- **Compensatory Equalization of Treatment:**
This type of social threat arises from the attitude of the researchers or external contributors. If, for example, teachers and parents perceive that there is some unfairness in the system, they might try to compensate, by giving extra tuition or access to better teaching resources. This can easily cause compensatory rivalry, too, if a teacher spurs on the control group to try harder and outdo the others.

These social effects are extremely difficult to minimize without creating other threats to [internal validity](#).

For example, using different schools is one idea, but this can lead to other internal validity issues, especially because the participant groups cannot be randomized. In reality, this is why most social research programs incorporate a variety of different methods and include more than one experiment, to establish the potential level of these threats and incorporate them into the interpretation of the data.

Cause and Effect - The Danger of Multiple Group Threats

Multiple group threats are a danger to causality caused by differences between two or more groups of participants. The main example of this is [selection bias](#), or assignment bias, where the two groups are assigned unevenly, perhaps leaving one group with a larger proportion of high achievers. This will skew the results and mask the effects of the entire experiment.

While there are other types of multiple group threat, they are all subtypes of selection bias and involve the two groups receiving different treatment. If the groups are selected from different socio-economic backgrounds, or one has a much better teacher, this can skew the results. Without going into too much detail, the only way to reduce the influence of multiple group threats is through [randomization](#), [matched pairs designs](#) or another assignment type.

As can be seen, establishing cause and effect is one of the most important factors in designing a robust research experiment. One of the best ways to learn about causality is through experience and analysis - every time you see some innovative research or findings in the media, think about what the results are trying to tell you and whether the researchers are justified in [drawing their conclusions](#).

This does not have to be restricted to 'hard' science, because political researchers are the worst habitual offenders. Archaeology, economics and market research are other areas where cause and effect is important, so should provide some excellent examples of how to establish cause and effect.

B. Establishing a Cause-Effect Relationship

<http://www.socialresearchmethods.net/kb/causeeff.php> 17 May 15

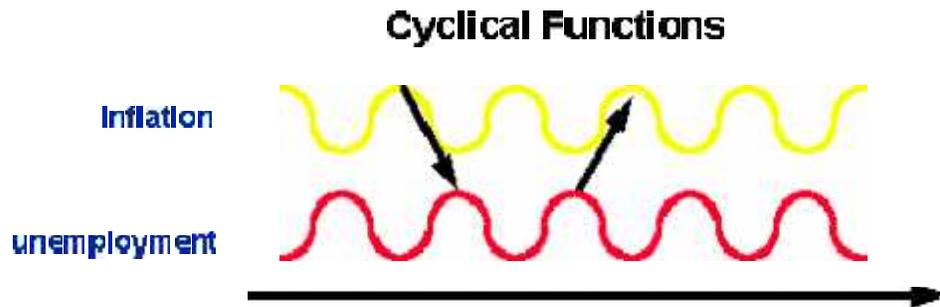
How do we establish a cause-effect (causal) relationship? What criteria do we have to meet? Generally, there are three criteria that you must meet before you can say that you have evidence for a causal relationship:

- **Temporal Precedence**

First, you have to be able to show that your cause happened *before* your effect. Sounds easy, huh? Of course my cause has to happen before the effect. Did you ever hear of an effect happening before its cause? Before we get lost in the logic here, consider a classic example from economics: does inflation cause unemployment? It certainly seems plausible that as inflation increases, more employers find that in order to meet costs they have to lay off employees. So it seems that

inflation could, at least partially, be a cause for unemployment. But both inflation and employment rates are occurring together on an ongoing basis. Is it

possible that fluctuations in employment can affect inflation? If we have an increase in the work force (i.e., lower unemployment) we may have more demand for goods, which would tend to drive up the prices (i.e., inflate them) at least until supply can catch up. So which is the cause and which the effect, inflation or unemployment? It turns out that in this kind of cyclical situation involving ongoing processes that interact that both may cause and, in turn, be affected by the other. This makes it very hard to establish a causal relationship in this situation.



- **Covariation of the Cause and Effect**

What does this mean? Before you can show that you have a *causal* relationship you have to show that you have some type of relationship. For instance, consider the syllogism:

if X then Y
if not X then not Y

If you observe that whenever X is present, Y is also present, and whenever X is absent, Y is too, then you have demonstrated that there is a relationship between X and Y. I don't know about you, but sometimes I find it's not easy to think about X's and Y's. Let's put this same syllogism in program evaluation terms:

if program then outcome
if not program then not outcome

Or, in colloquial terms: if you give a program you observe the outcome but if you don't give the program you don't observe the outcome. This provides evidence that the program and outcome are related. Notice, however, that this syllogism doesn't not provide evidence that the program caused the outcome -- perhaps there was some other factor present with the program that caused the outcome, rather than the program. The relationships described so far are rather simple binary relationships. Sometimes we want to know whether different amounts of the program lead to different amounts of the outcome -- a continuous relationship:

if more of the program then more of the outcome
if less of the program then less of the outcome

- **No Plausible Alternative Explanations**

Just because you show there's a relationship doesn't mean it's a causal one. It's possible that there is some other variable or factor that is causing the outcome. This is sometimes referred to as the "third variable" or "missing variable" problem and it's at the heart of the issue of internal validity. What are some of the possible plausible alternative explanations? Just go look at the threats to internal validity (see [single group threats](#), [multiple group threats](#) or [social threats](#)) -- each one describes a type of alternative explanation.

In order for you to argue that you have demonstrated internal validity -- that you have shown there's a causal relationship -- you have to "rule out" the plausible alternative explanations. How do you do that? One of the major ways is with your research design. Let's consider a simple single group threat to internal validity, a *history* threat. Let's assume you measure your program group before they start the program (to establish a baseline), you give them the program, and then you measure their performance afterwards in a posttest. You see a marked improvement in their performance which you would like to infer is caused by your program. One of the plausible alternative explanations is that you have a history threat -- it's not your program that caused the gain but some other specific historical event. For instance, it's not your anti-smoking campaign that caused the reduction in smoking but rather the Surgeon General's latest report that happened to be issued between the time you gave your pretest and posttest. How do you rule this out with your research design? One of the simplest ways would be to incorporate the use of a control group -- a group that is comparable to your program group with the only difference being that they didn't receive the program. But they did experience the Surgeon General's latest report. If you find that they didn't show a reduction in smoking even though they did experience the same Surgeon General report you have effectively "ruled out" the Surgeon General's report as a plausible alternative explanation for why you observed the smoking reduction.

In most applied social research that involves evaluating programs, temporal precedence is not a difficult criterion to meet because you administer the program before you measure effects. And, establishing covariation is relatively simple because you have some control over the program and can set things up so that you have some people who get it and some who don't (if X and if not X). Typically the most difficult criterion to meet is the third -- ruling out alternative explanations for the observed effect. That is why research design is such an important issue and why it is intimately linked to the idea of internal validity.

Chapter 36

Defensive reading of clinical literature: Handling the data

A. Power analysis - Did the study have enough subjects to determine an effect?

The bane of many clinical studies which fail to find a difference between two groups or which do not find a pre to post treatment effect is that they never had sufficient subjects to detect a difference of a reasonable size given the level of inter and/or intra-subject variability. In recent years the better clinical journals have been requiring authors to include power analyses of their data to show that they did have sufficient subjects to detect a clinically important difference if there was one. Unfortunately, most articles do not include such an analysis. Before you conclude that the intervention wasn't successful or that two techniques weren't different, conduct your own rough power analysis of the data. Look at the variability. If it is so great that it dwarfs the size of the difference being looked for and only a few subjects participated, find something else to read.

Many statistics programs have their power analysis sections set up so that you need only enter the information likely to be found in the results section of an article (such as means and standard deviations) to conduct a real power analysis in a few seconds without needing any of the raw data. Try it a few times and you are likely to be saddened by the number of articles which fail to meet even minimal subject number criteria.

B. Presentation of the raw data - the descriptive statistics:

1. Can you tell what actually happened during the study from the description of the raw data? You should be able to make your clinical decision from examining the raw data. The statistics should only be confirmatory if you are going to change your practice. If the result is so subtle that it can only be detected with inferential statistics, you need to question whether it is of true clinical importance.

a. Are the measurement / rating systems clear? If you can't tell why ratings were given, you can't trust the results.

b. Are there any clues as to what the distribution(s) of the major variables were like (frequencies, standard deviations, etc.)? If there aren't you actually have no idea what happened because variation could overwhelm the results.

c. Is enough raw data presented for you to get a feeling for what actually happened or do you have to count on inferential statistics. If you need to go by the statistics alone - watch out! The differences probably weren't consistently

clinically important enough to stand on their own.

2. Graphic presentation of the raw data: Watch out for misleading graphics. A graph or chart which doesn't present indicators of variation is frequently hiding something crucial. For example, the graph on the left of Figure 24 apparently indicates that pain intensity has decreased about fifty percent (from 6 to 3 on a scale of 0 - 10) from prior to after treatment. However, when an indication of variability is added, as is done on the graph on the right, it is apparent that the change is probably due to chance and regression to the mean.

Figure 24

Misleading amount of change when indicators of variation are not shown



No indicator of variability
Variability indicated by Standard Deviation

Relationships between events over time and between variables are especially distorted when one or more of the axes is a log function. Altman (Altman D: Statistics and ethics in medical research. Br. Med. J. 282: 44 - 47, 1981 cited by Spilker, 1986) gives an example of a very misleading graph purporting to show that immunizations resulted in a dramatic decrease in the death rate among children under 15 from diphtheria from 1871 to 1951. An approximation of this graph is reproduced in Figure 25 and clearly shows that the change in death rates is due to development of an antitoxin rather than the initiation of immunizations.

Graphs which do not have the origins for the abscissa and ordinate set the same or to 0,0 or which have breaks in either axis can be very misleading. For example, the graph in Figure 26 purports to show that test “a” costs about the same as test “b”. The graph on the left certainly

makes the two look very similar. Are they as similar as you are being led to believe? Can readers judge whether the difference is economically significant from a graphic depiction?

Figure 25 Distortion of data relationships by using a logarithmic axis

On the log scale, the change appears to



be when immunizations (IMM) are given while the change clearly came when the antitoxin was developed (ANTI) as shown on the arithmetic axis on the right.

LOG AXIS

ARITHMETIC AXIS

Figure 26 Distortion of visual interpretation by changing the origin of one axis



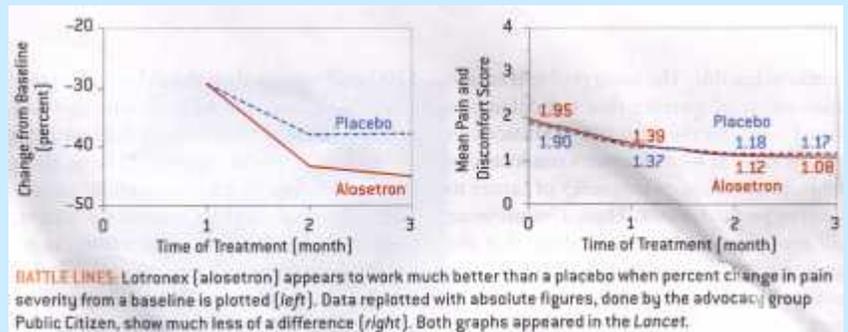
The real difference
The sales pitch

Some companies get especially creative when presenting their data. In the example illustrated in Figure 27, the only way to make it look like the placebo and treatment groups were significantly different was to not show the first month's data, to show changes in pain intensity as percent change rather than absolute numbers (this technique makes differences look larger), and, as you might expect, not to show any measures of variability.

Figure 27

Misleading information comparing a drug with a placebo:

- 1. First month left out (no difference bet drug and placebo)**
- 2. % change instead of absolute amount of pain**
- 3. No indication of variability**



From Scientific American 288 #2 Feb 2003, pages 15 - 16



C. Choice of statistical tests:

1. Justification: Given the experimental design and characteristics of the data, is this the test you would have chosen? If it is obvious that a Chi square is the way anybody would go but they used some unheard of test found who only knows where, something is probably wrong. If the authors didn't use the expected test, did they justify the use of the test they did use in an understandable way or is the explanation buried in statistical jargon. Nobody needs to hide behind statistical

gobbledegook if the results are clear and the study is well designed. As soon as you see statistical nonsense, take another look at the raw data. If the data aren't presented well enough for you to be able to tell what happened, it is probably time to stop reading the article.

2. Some questions to ask yourself include:

- a. What type of statistic is used? Are the analyses described appropriate for the purpose of the study and for the type of data?
- b. Is this the test anyone would have chosen for this design and type of data or is it some obscure test? If you never heard of it - watch out! They probably tried the usual one first and it didn't show what they wanted.
- c. Why did they use it? How does it fit in with the purpose of the study? . . . with the description, relationship, difference, effect, or prediction they wished to make?
- d. Every test has some element of chance connected with it. What is the probability that the results of the statistical test would be different with repeated sampling? -- usually expressed as "*p*"
- e. What is the meaning or clinical significance of the size of the relationship or extent of the difference?
- f. What are the possible explanations for the findings? Potentially confounding variables? Other internal validity issues that could have produced the results?
- g. Are negative and positive findings presented where appropriate?
- h. Are factors that might have influenced the results taken into account in the analyses?

D. Checking the analysis: You may not have to trust the inferential statistical analysis presented in the article. Just as is true for checking on the minimal number of subjects, very often you can redo much of the analysis yourself from the data usually available in the results section such as means and standard deviations. Most good statistical programs permit you to get good results for "t" and other similar tests by simply entering the means and standard deviations rather than all of the raw data. If the topic is important to you but you don't trust the analysis, it won't hurt to spend a few minutes entering a few numbers.

E. Relating the results to the discussion: Were the analyses sufficiently clear and clinically significant to warrant the conclusion? This is truly the key! All too many articles with weak results project firm conclusions calling for sweeping changes in practice.

Chapter 37

Pitfalls in study analysis

Examples of typical problems with experimental design and evaluation: (Very heavily modified from ideas in Principles and Practice of Research: Strategies for Surgical Investigators by H. Troidl et al; Springer-Verlag, New York, 1991.)

1. Sample size:

a. *"We tried our new fixator in 10 cases, with successful outcome in 8. Of 10 similar patients with similar conditions, the old fixator was successful in only half. Thus, our new fixator should replace the old one."*

This conclusion may be true, but only a few subjects were used with both instruments and there is only a small difference in the outcome (3 cases). This difference could be due to chance or be real. A "z" test of difference between proportions shows that the probability of the groups being different is only 0.35 which indicates essentially a random difference. It would take twenty-six subjects or so in each group to get a statistically significant difference at $p = 0.05$ given the difference in success rates.

b. *"We compared the effectiveness of temperature biofeedback and progressive muscle relaxation training for treatment of migraine headaches using a two group design with 20 patients in each group. The biofeedback group showed a mean reduction in headache frequency of 3.1 days per week (standard deviation = 2.8) while the relaxation group showed a decrease of 4.8 days per week (standard deviation = 3.6). The decreases were not significantly different so we conclude that the techniques produce similar results."*

The variability is so high that it would have taken 30 subjects per group to detect a statistically significant difference at $p = 0.05$. They simply didn't have enough subjects for the study to be meaningful. The authors could look at their data and see if one or two subjects caused most of the variability. If so, there might be a reason they responded differently from the others which would permit their being treated as potential outliers using the statistical techniques discussed earlier.

2. Common mistakes in analyzing studies:

a. **Incorrect use of multiple comparisons and Computer-Generated Significance:**
"Detailed surveys of health and economic status were sent to 500 patients treated at this facility for headaches over the last five years. Each survey provided 107 individual bits of information about the respondents. Each of these bits were correlated with whether or not the patient's headache was successfully treated. We were quite surprised at which items predicted success."

The authors used a significance level of 0.05. This translates to "the results could have happened by chance one out of twenty times". With over 100 variables, it would be surprising if there hadn't been about five associations by chance alone.

b. Not checking the data to see if it was similarly and normally distributed: *"We compared post-session fingertip temperatures before and after two weeks of temperature biofeedback training for 20 patients with Raynaud's syndrome. The pre-treatment readings averaged 53.8 with a standard deviation of 23.9 while the post treatment temperatures averaged 83.1 with a standard deviation of 61.4. Several of the final temperatures were much lower than would be indicated by the mean which accounts for the high standard deviation. A paired measures "t" test showed the temperatures to be significantly different so we conclude that the training was effective."*

The data are neither similar in distribution nor variation so do not meet the entrance criteria for the test. The test's result is meaningless. In fact, it is likely that some of the subjects learned the to warm and some didn't. These sub-groups should be factored out and the correct analytical techniques applied.

c. Follow-up data too short: *"We compared the occurrence rate of hip implant loosening among 130 patients receiving hydroxyapatite coated implants and 115 of those receiving identical but uncoated implants. After two years, both groups had statically similar failure rates so we conclude that the coating does not prevent failures."*

Two years is not in the range of when most failures take place. The follow-up is simply too short.

d. The conclusion isn't related to the data. *"We compared post-session fingertip temperatures before and after two weeks of temperature biofeedback training for 20 patients with Raynaud's syndrome. The pre-treatment readings averaged 63.8 with a standard deviation of 23.9 while the post treatment temperatures averaged 83.1 with a standard deviation of 61.4. A paired "t" test showed the temperatures to be significantly different so we conclude that the training was effective **in controlling Raynaud's syndrome among these patients.**"*

The data did not examine changes in Raynaud's symptoms so this conclusion can not be reached unless there is virtually certain knowledge that the change in fingertip temperature would produce this effect.

3. Common errors in use of statistical tools

(Modified and extended from (loosely based on) the "Handbook in Research & Evaluation by S. Isaac and W. Michael, 1971, Robert R. Knapp of San Diego)

a. Fail to carefully examine and plot the raw data as part of deciding which statistical tools to utilize and as the basis for understanding the results of the study.

b. Rely on statistics to determine decisions about value of the data instead of using them as a guide.

c. Select statistical tool that is not appropriate for proposed analysis.

d. Collect research data then try to find a statistical technique that can be used in the analysis.

e. Use only one statistical procedure when several can be applied to the data. This often leads to overlooking results that could have made a significant contribution to the results.

f. Uses statistical tools in situations in which the data grossly fail to meet the assumptions upon which the tools are based. This frequently results in the exact opposite answer from the way the data actually turned out and causes many clinical articles to be useless.

g. Overstate the importance of small differences that are statistically significant.

h. Use parametric statistics for non-parametric data.

i. Fail to set up a systematic routine for scoring and recording data.

j. Do not check scoring and data entry for errors.

Practical exercises for section D

1. What is the difference between descriptive and inferential statistics. Give an example of each.
2. Why should you “eye-ball” the data before doing an analysis?
3. What are three common distributions commonly occurring in psychological data?
4. Describe three common indications of central tendency and what are limitations of each?
5. What is the difference between parametric and non-parametric numbers. Give an example of each and explain the impact of the difference on data analysis.
6. What is power analysis, why is it important, and how does it work?
7. What is a “t” test and what is the impact of data distribution on it? When do you use it as opposed to a non-parametric test?
8. How do you choose whether to use a paired vs. independent test? Give an example of each set of data which would lead you to the choices.
9. What is an analysis of variance and what is it used for? Give examples of one and two way ANOVAs and explain the effect of repeated vs. non-repeated measures on the selection of an ANOVA design.
10. What is a meta-analysis and what is it used for?
11. What is a correlation and what is the effect of data distribution on its results? Include both parametric and non-parametric designs in your answers.
12. How do you analyze differences between (1) frequencies of occurrence and (2) proportions?
13. Discuss pattern analysis.
14. Discuss Survival / life table analysis.
15. Give an example of ways to graph data so the results are (a) misleading, and (b) accurate.
16. What is risk analysis? Make up a situation requiring risk analysis and make up a risk analysis table showing which choice you should take. Explain why.

17. Make up a data set similar to the one used in the inferential statistics sections which has three groups of at least six subjects each in which each subject is measured at least three times on one parametric and one non-parametric variable. Use your statistics package to go through every descriptive and inferential technique described in chapters 24, 28, 29, 30, 31, 32, and 33 to examine your data. Explain the results of each technique, describe its strengths and weaknesses relative to your data set. Be sure to explain the interaction effects in the two way analysis of variance and the impact of the co-variant in your analysis of covariance. When your data set does not support using one of the techniques, make up a special data set which meets the requirement. Keep subject numbers to a minimum.

Section E.

Synthesizing the elements to produce an effective study

Chapter 38

The protocol - incorporating statistics

A. Overview: In chapter three, the need to state the study's central questions in a testable way was emphasized. Hopefully, the study is structured so that each question you want to answer can be examined. By the time you get to the statistics section of the protocol, you (hopefully) already went through the process of weeding out questions for which you can't find a good design. If you realized that you had to drop so many important questions - or even the real question that interested you in the topic in the first place, you should have dropped the entire idea by now.

So, you are entering the statistics section with a good idea of the hypotheses to be tested and the design you will use to test them. Now you need to figure out which combination of statistical techniques are optimal for assisting in explaining the data you gather.

B. Major points to keep in mind:

1. There are no statistical techniques for some types of data. For example, there are no good non-parametric techniques for analysis of covariance or two way analysis of variance. **If your design will gather data which requires these non-existent techniques for analysis, you need to modify the design.** This is one of the most common pitfalls of study design and is why you should begin thinking about the analysis when you plan the design.

2. It is very unusual for only one inferential statistics approach to provide an optimal evaluation of any set of data. You could miss crucial information by using only one technique. For example, in the sample data set used in the last section, you would have missed important information if you only compared the intensity levels and frequencies using group comparison tests. At least as much would be learned by comparing the proportions of patients who did well in each group. Many apparently negative studies miss the fact the intervention really did have an impact, but it was restricted to a small sub-group of the patients. This group needs to be noticed and then identified.

3. Plan your statistics so they will help you defend against spurious relationships. Even if you avoided using the shot-gun approach in gathering data, whenever you compare more than a few variables you are exposed to the possibility of having spurious relationships show up. This means adjusting your expectations in light of your knowledge of the disorder, your subjects, and the variables you measured. This "Bayesian" approach is crucial to avoiding serious and silly mistakes in your conclusions.

C. Aligning data collection and recording methods with the statistical package:

1. The statistics package you will use requires the data to be set out in particular ways for each test. You need to arrange your data sheets in such a way that the data will already be in the format and arrangement required by the statistics package so you don't have to reenter or reformat your data.

2. By determining which variables are required for which tests, you may realize that some of the variables are not really necessary for you to reach your conclusions.

D. Reevaluate variables which appear to co-vary: If you have performed a pilot study, look for variables which have co-varied with each other so much that there is no need to gather all of them as only the one that has the most relationship to the information you want is really needed.

E. Aligning the statistical methods with the hypotheses and the methods: The best approach is to sit back and think about what you really want to find out. To answer each of your questions, you are going to have look at a different sub-set of the data or look at it in a different way.

1. First, chose the overall descriptive methods you will use to look at the raw data. This is crucial as you need to be able to identify unanticipated asymmetries, sub-groups, and outliers

(not to mention differing variances) before stuffing the data into inferential statistical formulae which may no longer be appropriate for them.

2. Second, look at each hypothesis and the design used to elucidate it. Choose several inferential techniques which are most likely to be useful with the type of data to be gathered and think about whether the types of answers they give will really answer your question. Suggest several approaches to answering each hypothesis in your protocol.

3. Think about the difficulties which could arise if the data are somewhat different than you anticipate and suggest several approaches which will still permit you to approach answering your question.

F. Explain and support your choice of inferential statistical techniques: Most people really don't know what the strengths and limitations of inferential statistical techniques. Take a short paragraph to explain why each test will tell you what you are trying to find out and why the type of data you predict are appropriate for that test. Unless you are inventing a new statistical technique, you got it from someplace. Cite any unusual techniques so reviewers can look up what they are used for just as you cite supporting evidence for your hypothesis in your introduction. This practice helps give reviewers confidence that the investigators have really checked out the technique.

G. Test your statistical package: If you have done a pilot study or have data available which approximate what you are likely to gather, try out your statistical package to be sure it can really handle the techniques you want to use. You may get an unpleasant surprise when you find out that it won't graph out your data in a way you need or prodding it to do one of the analyses may be either ludicrously difficult or require information you weren't going to gather or can't estimate.

Sample study - aligning the statistical approach with the hypotheses and design

1. For Sample 2 (the controlled headache sample study):

Hypothesis "A" lines up with analysis technique 1" while hypothesis "B" lines up with analysis technique 2" so both the investigator and the reviewers can tell exactly how they relate to each other.

How could the investigators have done a better job describing the overall descriptive approach to the data?

2. For Sample 3 (the evaluation of mulehailer's syndrome): The hypothesis is not stated in a testable way so it is probably just as well that the statistics section got left out.

Chapter 39

The grant - extramural funding process

A. Why grants?: Research is expensive! Even if you are going to run the study yourself for free (and, thus, do not need a research associate), you still have to pay for supplies and perhaps some equipment. Depending on where you are doing the research, you may have to pay overhead for use of the space. You may even have to pay the subjects.

Very few people have sufficient independent wealth which they wish to use for their research so the money has to come from some other source. The days when research oriented academic departments and institutions could give students and faculty funds to conduct a major project are just about gone. A few still have funds for start-up pilot projects but these tend to be very competitive and/or limited to new faculty. Thus, if you are going to do research, you have to find somebody willing to support it.

B. Sources of funds:

1. Funding from government agencies: Funding for general clinical research and training is available from the National Institutes of Health (NIH). The military and a few other government groups fund a very few, highly specialized projects in support of their very limited needs. NIH gives grants and contracts for various types of research performed by various classifications of people. NIH is very choosy about whom they give their limited funds to. You almost have to have the correct pedigree. Being a physician doesn't help much but it helps if you also have a long list of peer reviewed publications in the area to which you are applying. Don't waste your time applying for an NIH grant for other than start-up or new investigator programs unless you are part of a very strong, well established research team. Even then, unless you already have NIH grants, it is very unlikely that you will be funded. NIH does not like to risk their funds so they only support research that is almost sure to work. This means you must have completed so much pilot work that you almost do not have to perform the study. Many representatives from NIH have publicly admitted that they know that they are supporting work that has actually already been completed so the investigators can use the funds to do pilot work for future applications. They almost never fund any project that (a) has new ideas or (b) who's theoretical basis the review committee does not agree with regardless of pilot data or scientific merit.

Be prepared to spend a minimum of 20 full time days preparing the grant. It must be printed with quality at least equal to that of a laser printer and has to be gorgeously formatted. If

you can't do this, don't waste your time because they won't take you seriously regardless of the scientific merits of the project.

2. Private foundations: The non-profit sector should not be ignored. Most, but not all, of their grants are smaller than NIH grants but they are much easier to get as their paperwork is far less onerous and they are under less pressure to support a wide variety of projects. Several computer bases (such as the Foundation Center) and numerous books list private granting agencies and the types of research they support. Several of these are listed at the end of the chapter. As long as your project meets the EXACT needs of the foundation, you should not hesitate to apply. My research has been (and is) funded by several non-profit organizations. I have found them very willing to discuss what they are interested in and to keep a good relationship going.

3. Industry: The "for-profit" sector is much maligned as a source of funds because organizations which want to make money from your work might influence you to influence the results (even unconsciously) through study design factors and interpretation of the data. However, industry does need to find new products if they are going to remain ahead of the competition and they do need to prove that their products work - and work better than the competitors' products. Thus, industry is frequently very willing to support research in areas they are interested in. Many large corporations have separate research groups and foundations to handle applications for funding. Even in this day of very tight money, sales personnel visiting health care facilities or their supervisors frequently have discretionary funds they can use to support start-up projects which sound interesting. If you can make an arrangement with a company which suits your ethical standards and doesn't conflict with your organization's requirements, there is no reason not to pursue this route for funding. I work with several for profit groups without experiencing any pressure to do anything but excellent research incorporating their products. One of the companies is very interested in design changes so I work quite closely with them on this aspect of the study. The other is interested in how well their product performs with a unique population for whom I need to find a treatment for a severe problem. This has resulted in a senior scientist from the company and I becoming close colleagues and co-investigators as we each use each other's knowledge base to build a better project. These interactions are not unusual when working with industry as they have their own scientists working toward their own goals. However, the interactions are frequently quite impersonal and little different than getting a grant from the government except that they require relatively little paperwork. Of course, I make sure that my institution handles the funds and takes care of all of the legal requirements for the relationship.

C. Grantsmanship:

1. Have an excellent idea which is demonstrable with one good project: Don't apply for a grant unless you have a really good idea which can be investigated with a traditional study design to produce an important answer. If your idea isn't good enough to be considered important, the odds of its being funded are quite low. If the idea is good but there isn't (a) enough supporting or pilot data and (b) a well worked out methodology to insure high odds of your successfully completing the project, it is not likely that someone will take a risk on funding it. This is because there are far more good projects around than there are funds to support. Thus,

you need to convince the granting agency that the odds of your finding out something important that they care about are very good. It is up to you to convince the reviewers that you and your team are the optimal group to get the project done and that the project is not only important to the world but doable.

2. Apply to the folk most likely to fund you and your work: Before you apply for funds from any group, be it a foundation, a branch of the government, or industry, call the agency and talk with people there in sufficient detail to find out *what topics* and *who* they actually fund. If they don't fund your type of organization or people with your academic background (e.g. they only fund neurologists), you need not waste time applying unless somebody high in the decision making process tells you they are willing to make a specific exception in your case. I do not mean that somebody says, "go ahead and submit - see what happens". Many organizations will fund equipment but not support personnel salaries and vice versa. Some will not pay your salary but will pay for everything else. Make sure you know what they won't do so you don't apply for something that will result in an automatic refusal unless you get an actual exception. I was fortunate enough to have gotten such an exception but it took a very careful explanation backed up with considerable facts.

3. Find out what the granting agency wants to see: Don't start writing until you get a copy of several grants the organization has actually funded recently. Some organizations will not send a copy but they will send a list of who received their grants. You call a few of these people and ask them to send you a partial copy. This is crucial because all organizations have a culture of what they look for when they read a grant. This is especially true of NIH. If you don't have the correct buzz-words and concepts of the day, you won't get funded. Try to match the style, depth, and organization of the grants which were funded.

4. Find out who is going to review the grant and review their publications: Many non-government and all NIH and VA organizations provide a list of reviewers upon request. It is usually a bad idea to contradict the work of someone on the review panel without a very careful explanation of why. People really do turn down grants because of intellectual disagreements. They are also very likely to turn down your grant if your methodology or intervention is not of a type they like regardless of pilot data or logic. Approaches which are perceived to be either "complimentary" or alternative medicine have little chance with NIH unless you are very careful and very lucky. At this time, NIH's Office of Complimentary and Alternative Medicine is a reasonable place to apply. They have very helpful advisors who can help you through the administrative process.

If your grant is rejected but gets a score which indicates that the agency thinks there is some merit to it, you can make the changes recommended by the reviewers and resubmit it. In the case of groups such as NIH, this is nearly the norm. If you feel that the reviewers were ignorant, prejudiced, missed the idea, etc., you can provide a detailed justification requesting different reviewers or consideration by a different review panel. Don't expect to do any better with the new group.

5. Provide the proof you can do the work: Your own and the members of your team's research track record with the granting agency and other granting agencies means at least as much to the reviewers as the project for which you are requesting funds. You must get your co-

investigators and consultants in line with the kinds of people the granting agency expects to see. Few groups are going to give money to a novice who has never done any research in the field. Arrange your supporting people so that it is obvious that the required expertise and experience is actually available (as opposed to lip service) for every aspect of the work you can not do yourself. Your first shot might be as a co-investigator with (a) a very well published, senior person or (b) the profession of the person the organization will give money to as the principal investigator even though you do all the work.

6. Write it superbly: The actual writing of the grant has to be done very, VERY well. The reviewers will have many grants to read and they are going to ignore yours if it is not superbly written. You are trying to sell an idea to people who have never heard of it before. Your writing has to be clear, concise, and, above all, convincing. That means providing supporting material which leads the reviewers to predict what your project will be because it is the logical next step in an important chain leading to a crucial discovery.

The reviewers get the main idea and make their first emotional decision about your project from the title. It has to win them to your side as well as explain just what you want to do.

The most carefully read and best remembered part of the grant is the summary. Regardless of where it is in the format of the project, the reviewers will read it first to get an overview of what you are trying to accomplish and why. If you don't sell the project with the summary, you aren't going to sell it. Thus, you need to spend a very disproportionate amount of time producing a very clear, concise summary.

7. Do not apply for a government grant without reading a book on grantsmanship. You may also wish to read such books to increase the likelihood you will write a grant good enough to win funding from a private or non-profit group. These books usually give good advice on where to apply.

D. Taking the money: Most, but not all, grants are made to institutions rather than the investigator. The institution normally owns the equipment purchased or given as part of a grant and administers the funds. The principal investigator has to be very careful about how the funds are accepted and traced.

If you work for the Federal government, DO NOT under any circumstance, accept anything from anybody not part of the Federal government regardless of what they tell you!!! You must not accept the loan or gift of equipment, supplies, support personnel etc. You must not accept money from any source other than the Federal government. Everything coming into a Federal installation for research has to be handled in complex, cumbersome, time consuming, frequently expensive and wasteful ways before you can touch it. Note that the regulations change very frequently and nobody seems to agree on what they are. Every institution seems to apply the regulations differently and every administrator gets very, very worried when having to actually approve accepting anything from outside because neither the legal system nor higher headquarters give consistent support or advice. If a person in the Federal government - either a member of the military or a civilian employee, who is not specifically authorized (in writing as part of their job description) to accept and spend money does so, very severe penalties can ensue and have done so. Hapless clinicians have destroyed their careers on the rocks of ignoring this prohibition!

Regardless of what kind of organization you work for, get written authorization from the administration to accept the grant - even if it is just the loan of a piece of equipment. If it breaks or disappears you may wind up paying the bill unless you have such authorization.

If you are in private practice, don't take anything without a very clear, written document signed by both you and the grantor specifying all the conditions of the grant including return, repair, security, what you need to provide, etc.

When you accept a grant, be sure you do not give up your ownership of the data and the right to publish whatever you find. Many for-profit organizations would like you to agree not to publish findings they do not approve. Do you really want to take their money with that kind of restriction?

E. Sources of information about grantsmanship and sources of grants: Your best bet is to *locate agencies on the Internet/www*. Virtually every major granting group, including NIH, has very clear, up to date web sites which are very easy to find. Only if you can't locate the web sight, look for:

The Foundation Center: National directory of grant-making public charities. The Foundation Center; 79 fifth Ave; NY NY 10003.

Research Grant Guides, Inc.: Directory of health grants. Research Grant Guides, Inc. PO BOX 1214; Loxahatchee, FL 33470.

Public Health Service: Omnibus solicitation of the PHS for small business innovation research grant and cooperative agreement applications. PHS95-3 (407) 791-0720.

Division of Research Grants, NIH: Preparing a research grant application. (301) 435-0714; Division of Research Grants, NIH, Bethesda, MD 20892.

Arthritis Foundation: Research Awards Directory. (404) 872-7100; Research Department 1314 Spring St, NW; Atlanta, GA 30309

National Association of Orthopaedic Nurses: Grant Resources Guide. (609) 256-2310; NAON; East Holly Ave, Box 56; Pitman, NJ 08071.

Ogden T: Research Proposals: A guide to success.

Troidl H: Principles and Practice of Research: Strategies for Surgical Research Springer-Verlag, New York, 1991.

Chapter 40

Writing and presenting the study

A. The problem: Doing an excellent study requires an enormous investment in time, energy, and ego. All too often, much of that tremendous effort is wasted because the results never get to the people who need to know about them. This is usually because the results are presented so poorly either at a meeting or in an article that nobody notices, remembers, or understands them. It takes a real effort to present well regardless of whether the presentation is in the form of an article, an oral presentation, or a poster. If you haven't presented before or aren't a particularly good writer, you might want to read Debora St. James' (1996) book on writing and speaking for excellence or the section on writing and presenting in Bordens and Abbott's (1991) research design book.

B. Strategy for the article:

1. Getting your foot in the door: According to Current Contents, few articles are read beyond their titles. Those few get to have their abstracts scanned and perhaps read in detail if the scan holds the reader's interest. The rest of the paper is virtually only read if the reader has some specific need for the information. For example, the reader may be a student preparing for a journal club or an investigator working in that or an associated area. These people tend to concentrate on the result section, then the conclusions, and finally the methods. They may never really look at the introduction at all or may just breeze through it.

With this in mind, you need to develop a strategy which will enhance the odds of getting your point across to as many people as possible. If most people are only going to read the title, the point has to be encapsulated in the title with support in the abstract. Of course, sometimes people abuse this strategy by making the title more sensational than the results warrant. Occasionally the attempt to get the whole story into the title results in it being so long and convoluted that nobody grasps its import on initial scan and the best chance of getting the point across is lost.

2. Write for your audience: If you are writing a highly technical article for a highly specialized journal and don't care if anybody outside the tiny sub-sub-sub specialty ever uses your work, you can have a technical title and abstract with lots of vocabulary intelligible only to your few cronies who read the journal. This really isn't a good idea even with a technical journal because people may not be able to grasp the concept you want to get across when you string several fifty cent words together. This is a very common problem in surgical specialty journals because the authors assume that all the readers will be other surgeons from that specialty and use specialized jargon in the title and abstract. This prevents other people who might want to use the information for parallel work in their own specialty from even realizing your work is related to

their interest because they'll never get past the title. Many surgically and behaviorally oriented journals seem to be full of acronyms for concepts and phrases which nobody outside their areas know. A scattering of these in the abstract induces just about everybody not in the clique to find something easier to read.

If you are writing for a general journal keep the jargon out of the title and the summary/abstract. In general, if you write for the college educated, intelligent lay-person people will understand your work and you will force yourself to write clearly enough to really make your points. Even the "experts" will get a clearer picture of what you did because they tend to ignore segments that look like standard jargon-bites. The article won't be much longer. I have never seen a paper rejected because it is a few dozen words longer due to clear writing.

3. Key words: The vast majority of people doing research or trying to keep up with a field read only one, or possibly two, journals. These tend to be very sub-specialized so breaking news in related fields is frequently missed unless it makes the TV news or the equivalent. The situation is complicated by the plethora of journals covering overlapping fields so an article exactly in a clinician's main area of interest might appear in any of several dozen journals. The result is that, when clinicians want to catch up on what is happening in a field, a literature search is performed.

If your selection of key words doesn't match the search terms, few people are going to find your article. Spend time selecting those words! Select words that are general enough so somebody not familiar with your exact work but interested in the topic is likely to use them. Check that they produce the articles you think they would by using them in a search of your own to see what comes up.

C. Make it clear or you are wasting your time: The medical literature is rife with articles which can not actually be read or interpreted. They are frequently so poorly written that they do not get their points across. It takes a great deal of time to write a clear, concise article. Once you write the first draft, look at it with a very critical eye toward shortening and clarifying it. Have several of your colleagues review it for both content and clarity. It also needs to be reviewed by an "English major" type of person who is **not** particularly familiar with your field in order to get the rough parts out. If they can understand it, your readers probably can also. You do not have to take all of the advice and can leave in technical aspects you are certain that your readers will understand.

D. Where do you start?: If you already wrote an excellent protocol, you have done most of the work. The protocol summary forms the first part of the abstract. You should have incorporated the results of an exhaustive literature search into the protocol's introduction and, hopefully, synthesized the information while creating the argument for why your study should be done. This introduction can be cut down to form the article's introduction. Remember that much of the synthesizing you did will appear in the article's discussion section when you relate your findings to those of the field. Obviously, the protocol's method section will translate to the article's section of a similar name. That just leaves the result section for you to write from scratch.

E. Optimizing the sections:

1. The title: As noted above, the title is crucial to getting your point across since this is the only part of your effort most people will ever see unless they are attracted to read more. The title has to be short and clear enough to be quickly grasped yet tell the story. Convoluted phrases with technical jargon don't get read because of the difficulty in translating them.

Contrast the following titles:

a. Successful treatment of migraines with pulsing electromagnetic fields: Results of a double-blind, placebo controlled study.

b. A double blind, placebo controlled study of the treatment of common and classic migraine headaches using PEMF.

Title "a" starts by letting you know that something worked so perks your interest. It follows with what was treated and tells you what the treatment was. Finally, it tells you what kind of study it was. If this last bit of information hadn't been added, you might skip reading further because it would be natural to assume that this was another in the plethora of uncontrolled single group conglomerations (clinical replication trials) which glut the journals with claims of being able to cure every case of whatever they were working on but which are never seen in the literature again unless an abject failure is reported. Title "b" starts with the design so you might never get far enough to tell what they were doing because all the extra words (such as "common" and "classic") slow you down. Virtually nobody except a few experts know what "PEMF" stands for so using the acronym is more of a negative than an explanatory attractant.

2. The abstract: If people get sucked in by the title, they will scan your abstract. This is your big chance to give them enough details to get the gist of what you did. If they want all the details, they can read the article. Keep this under a hundred words if you are reporting a simple study. If the study has multiple parts so needs to be longer, set the abstract up so everything the readers need to know is in the first fifty words or so because they will be scanning more broadly the further down the abstract they get. Put the punch lines right where you would expect to find them if you were scanning an article you weren't sure you cared about.

Clinicians are likely to only be looking for the highlights of how many of what kinds of patients you used, what the basic design was, what happened, and what this means to their clinical practices. Yes, you need to say how many males and females you had, what the age group was, and some of the statistics supporting the outcomes. However, you can simply give the number of males and females and the mean of the overall group's ages along with the group standard deviation rather than spending two lines on details of the subjects nobody needs to decide what happened in general. You can also leave the diagnosis and categories with something such as "20% of the participants had classical migraines with the remainder having common migraines. People with mixed and other types of headaches were excluded." You can leave the actual diagnostic criteria, frequency of headaches, etc. for the full paper.

The results can give the starting and ending means and standard deviations along with some statistic showing that the change was statistically significant but readers will be making their judgement based on changes in clinical severity.

Contrast these two ways of giving the highlights of the results:

- a. 31% of the subjects showed at least a 50% decrease in headache activity while 15% were headache free for at least two months.
- b. Headache activity (a conglomerate of frequency, intensity and duration) changed from 6.2 +/- 4.9 during the month before treatment to 3.7 +/- 2.1 during the month afterwards which is significantly different at $p = 0.023$ ("t" = 2.35 with 48 DF).

Type "a" catches people's attention more than "b" because it is more intuitively meaningful to a clinician who has to treat migraines. I am not saying not to put the statistics into the abstract. Rather, make sure you begin with something that makes intuitive sense.

3. The introduction: This has to be short and convince readers that the study had to be done and that your approach was the one they would have used. The problem here is balancing depth with brevity. If you leave out key work, people - especially the reviewers - will notice. If you leave out work that contradicts your hypothesis, people will get annoyed and your credibility will be so low that knowledgeable readers may not trust your work.

The introduction also has the role of justifying and explaining the measurement techniques you are using. If people don't understand what the techniques measure and the relationship between their measurements and the clinical variable, they aren't going to trust the work.

4. The method section: Any reader with the correct mix of skills and access to the equipment and subjects you described should be able to replicate your study exactly from reading the method section. If you used some standard technique, you must not only cite where the details are published but give a brief description of how you used it and any deviations from the standard technique. The design has to be very clear. Many journals encourage diagrams of key equipment which is unfamiliar to most of their readers as well as reproduction of unique logs, pain scales, etc.

This is where you explain and justify your statistical approach. You should assume that your readers know little about inferential statistics and truly let them in on what you were doing. If you used an unusual test or used a different test than the one most knowledgeable readers would have used for that design, be very clear about why or people will assume that you tried the regular tests and didn't get the results you wanted so hunted out a test that did.

5. The result section: Take the time to do this well. Most people want to see enough of the raw data to make up their own minds about what happened so have at least one table with as many of the individual subject's data points as possible. Simple, uncluttered graphs can tell a much more effective story than a complex table or an endless mass of mind glazing statistics mixed into convoluted text.

Try to put most of the descriptive and inferential statistics into tables. Few journals will reject a good graph when it actually helps explain what happened but they dislike both (a) overly simple graphs whose information is just as easily understood in one sentence and (b) complex

graphs with numerous lines representing complex breakouts of groups which don't really help the reader to reach a conclusion about what happened.

The result section is not a good place to discuss problems but you do need enough verbiage for readers to understand what the results were.

Your graphs must be professional quality. The days when you could send something in with tape stuck to graph paper ended with the personal computer's entrance into graphics. Every good word processor can make graphs at least as professional as anything most medical illustrators could ever do at their best. Take the time to do it right!

6. The discussion / conclusion section: This is probably the most often botched section in an article. The author has at least three main points to cover.

First, of the results must be explained and clarified. The significance of unexpected sub-groups, etc., need to be explained and problems with the study need to be discussed in relation to their impact on the overall results.

Second, the results need to be incorporated into a real synthesis of the literature reviewed in the introduction. This is the part that tends to be ignored. If everybody else had lousy results with techniques similar to yours but your study shined, you should explain why yours might be different. This is the place to note that all of your subjects were young and otherwise healthy athletes while everybody else using the technique studied unconditioned, ancient folk with multiple medical problems.

Third, you need to explain where your results fit into the decision to change clinical practice. If this was a basic study, you need to indicate what comes next to move in the direction of clinical implications. If it was a clinical study, you really need to put the results into perspective with current practice. For example, if you have a great new treatment for migraines but only tested it on fifty subjects in two studies, do you want people to drop everything and adopt your treatment, do you want them to start lots of clinical trials on the idea, or do you want to let them know that large, long term effectiveness/safety studies are the next step?

7. The authors: I left this for last because it is more of a political point than a structural one. When you write a good paper that is likely to get published, suddenly everybody wants to be a co-author. You need to make sure that everybody who actually helped you perform the study gets included as a co-author if they weren't paid specifically to provide the assistance. The days when the department chair was automatically a co-author are dying but not dead yet. Powerful figures still force their way onto publications under various pretexts. Various professional organizations have given deep consideration to who deserves to be an author and have reached essentially the same conclusion espoused by the "Uniform requirements for manuscripts submitted to biomedical journals", which state that substantial contribution to **each** of the following are required to be included as an author:

- a. Conception and design, or analysis and interpretation of the data.
- b. Drafting the article or revising it critically for important intellectual content.
- c. Final approval of the version to be published.

If you were paid to perform your role in the study and do not exceed that role, you would normally not be a co-author. Thus, technicians, biostatisticians, etc. are not co-authors unless they do something beyond their paid role.

Many journals now require each author to sign a statement that they have met the above three requirements. It is fraud when a chairperson or other power broker forces his/her way onto

an article and signs such a statement.

F. The oral presentation: Most oral presentations of individual studies presented at scientific meetings are ten to twenty minutes long with fifteen minutes being the most common. It is very difficult to get and then keep your audience's attention long enough for them to get and then be convinced of your point. Your presentation has to be very clear and concise with a few very well selected, absolutely clear visual aids. Unless you happen to be the first speaker, your audience is already fatigued from hearing other talks and the competition from extraneous sounds and the slow pace of talking vs. thinking has probably left much of the audience daydreaming about more interesting topics. You will lose your audience after just one moment of droning on semi-intelligibly about numbers in front of an unreadable slide covered with unintelligible lines or a table stuffed with even more numbers.

Some pointers include:

1. When you speak, make sure that your topic is obviously relevant to the audience's clinical interests or they won't begin to listen.
2. Decide just what message you want to get across then plan your talk and slides based on how you are going to do it.
3. Tell the audience what you are going to tell them, tell it to them, and tell them what you told them - then ask for questions.
4. People have a lot of trouble following oral presentations because they can't flip back a page or two to pick up a point they missed so you need to be very well organized and leave out all information which doesn't lead directly to your goal.
5. Your slides need to be very clear. Nobody can follow a graph with more than a few lines and tables with more than three columns and rows (or so) are hopeless because of size and complexity. Having the main points of the talk on successive slides presented as you go is a big help but keep the verbiage down to a few lines per slide. Get rid of any slides the audience doesn't need to follow your talk. They are simply distractions. Slides full of text can't be read at the same time you talk so people miss both.
6. Your key slides are the title of your talk, a simplified diagram of the study design, a summary of the key results, and a few slides on the conclusions.
7. Practice your talk so you get the length right. Assume that you will go slower on stage. While practicing, get some emotion into your voice (other than terror) and learn to give the talk from an outline. People that read their talks well are incredibly rare. Usually they present in rushed, boring monotones which few people follow.

G. The poster presentation: Posters are the medium by which most trial and pilot results are presented. This is usually the novice researcher's first exposure to the scientific public. Posters take considerable care to do well because of the need to balance attractiveness - so people will stop and look - with the need to get enough information across in less than a minute to demonstrate your point. People drift through poster sessions and are attracted as much by a really good looking display as a clear, relevant title. This is the same issue as faced by people writing an article. If you type out your poster on computer paper and pin up a series of sheets covered with solid text without a huge title board, the odds of anyone stopping by are very low.

A few pointers:

1. Just adding a prominent title board on a bit of cardboard helps tremendously because all us middle aged, visually impaired scientists can make it out from more than two feet away.

2. The viewers will get your point from the title and the summary or miss it entirely. The summary should be less than a hundred words and be typed out in very large, dark

text. Nothing less than 40 point will do for

the summary because nobody can read 12 point type such as this from more than two feet away and people aren't going to put their noses against your poster unless they are certain there is something on it they absolutely need to know. If they want more information than is in the summary, they will look at the rest of the poster.

3. Include a very clear diagram of the experimental design and a summary of the subject's characteristics. Several excellent graphs and tables no more complex than you would use for a slide should cover the results.

4. Keep text to a minimum. Nobody wants to spend five minutes in front of your board trying to read an article.

5. Provide handouts with the full manuscript so people don't feel that their only chance to get the information is from what little they can glean from the poster. Handouts insure that people can use your information if they want to because they have it. It also gives you a second chance to get your point across because they'll see your title again when they sort through all the stuff they picked up at the meeting prior to dumping 99% of it.

6. Posters are the greatest networking tool for junior investigators I have ever heard of. Although it is true that a few senior folk stop by and pontificate endlessly to no apparent point, you have the golden opportunity to personally convince many people of your point and to learn an enormous amount from other scientists who can help make your next project stronger. This is an opportunity to form tentative collaborative efforts and learn who else is working in your area. Really senior people with whom you have absolutely no hope of getting an appointment are very likely to stop by and chat with you. They can guide you to sources of information which you would never know about or simply are not available without an entre from a senior person. They are also likely you remember you at least hazily so when you want more information, you have somebody you can call who can actually help.

Chapter 41

The publication submission and review process

A. Choosing a journal: This can be a difficult but crucial decision as you could submit the same manuscript to five journals which cover your topic area and get five different responses from outright rejection through various levels of recommended changes to direct acceptance. Your work may never be published if you make the wrong choice several times in a row.

While there are a few well known journals covering major areas of health care, there are also many thousands of clinical journals which cover an astonishing variety of tiny slices of the clinical realm. Many are of the typical peer reviewed variety which cover a significant segment of a field and may be put out in conjunction with that segment's professional organization. Typical examples are the Journal of Bone and Joint Surgery and the Clinical Journal of Pain which have wide readerships beyond the members of their supporting organizations. Others may represent a vanishingly small special interest group such as the Association for Applied Psychophysiology and Biofeedback's Journal of Applied Psychophysiology and Biofeedback. Journals may be put out directly by publishers attempting to attract an audience interested in a particular area such as complimentary medicine.

Most journals are interested in receiving manuscripts within their scope of interest but some are by invitation only.

You need to be sufficiently familiar with the literature in the area you are researching to get a feeling for which journals recently published studies similar to yours in both subject matter and type of design. The "recent" part is important because editorial policies tend to change every few years as editors turn over. You may have done a terrific study which shows that exposure to green cheese cures warts but sending the manuscript to a journal which never publishes alternative medicine treatments may be a waste of your time unless you have some very concrete reason to think they will accept your effort.

You need to optimize your odds of getting your manuscript accepted by making an informed decision. The information to decide whether to submit to a particular journal comes from such sources as:

1. Your own and your colleagues prior experience publishing in that journal.
2. A review of the journal's recent articles. NOTE that this review needs to include the status of the published article's authors and importance of the findings to the journal's target audience. If they don't take studies that make tiny advances which may lead to an important advance after a few more studies, don't bother to send such manuscripts. The articles need to be inclusive not only of the clinical topic but the specialty and type of technique your manuscript incorporates. Surgical journals rarely accept articles which solely discuss medicinal interventions for problems treated both surgically and medicinally.

3. Phone discussions with the editor about the manuscript. I have done this numerous times with great success both in saving myself lots of time submitting an article doomed to nearly certain rejection and in paving the way for acceptance by helping the editor understand what was coming.

4. The journal's reputation in your field. Frankly, there is little point to submitting a mediocre article to a high quality journal such as the Journal of Bone and Joint Surgery.

5. The journal's overall breadth of coverage and status: Journals such as Pain are high quality and take manuscripts on a wide variety of pain related topics but are not likely to accept a manuscript which would not be of interest to a wide variety of readers. If your study covers a very specialized area and does not demonstrate a global advance in understanding or treatment of pain, find a more specialized journal. Don't bother with the New England Journal of Medicine, Science, etc. unless you have a solid reputation and your study truly represents an important advance in the understanding or provision of health care.

6. What you think of your status and reputation: If you are very well known in your sub-specialty and have published numerous times in a particular journal which represents that sub-specialty, you have a much better chance of getting a study which is slightly off the journal's center of mass accepted than people who don't have that track record. If you are a newcomer to the field and don't have a track record, seek journals whose interests are right down the alley of your manuscript.

B. Following the journal's instructions for manuscript preparation and submission:

Authors who ignore the usually very clear "instructions to authors" causes enormous frustration for both the authors and editors who have to return manuscripts unread. Most journals have a very clear policy on the types of manuscripts they are interested in - both in subject matter and the way it is handled. They specify the style and appropriate length for case reports, reviews, investigative studies, etc.

The editors really do want you to use the margins, type faces, citation formats, etc. that they specify and have usually instructed their editorial assistants to return manuscripts which deviate significantly from the instructions. Your best bet is to read the instructions carefully and do what they want. Be especially careful to follow the way the journal wants references cited in the text and how they want the manuscript submitted. If they want three copies with separate originals of art work - just do it.

Many journals now prefer electronic submissions or a disk accompanying the manuscript. If they want a disk, pay attention to the format they request.

C. The initial review process:

1. Your manuscript will usually be looked over by an administrative assistant to insure

that it meets minimal requirements and then either be returned for correction or forwarded to the editor. The administrator usually sends some form of communication indicating that the manuscript arrived and has been assigned some review number.

2. The editor will take a brief glance at the manuscript and decide whether to send it back to you or forward it to a specialty sub-editor for further review. In some special cases, the editor may perform the initial review.

3. If the journal is of the "peer review" variety, the editor or sub-editor will pick between one and four people who the editor feels have expertise in the general area covered by the manuscript. If the manuscript covers a clinical technique not normally used for a particular disorder, the editor will try to find people who are experts in the technique and other people who are experts in the disorder. The pool of reviewers for the journal may consist largely of members of the journal's editorial board with outside experts brought in when necessary. If there is a scientist who you feel should not review your manuscript because they are a competitor who may steal your idea or you know to have a viewpoint so opposed to yours that they would reject the manuscript out of hand, you can request that this person not be a reviewer. However, you can't normally suggest who should review it unless the editor asks. Such requests are vanishingly rare.

Although editors usually give reviewers only four to six weeks to return a review they have agreed to perform, it can take many months to get reviews back from busy clinical scientists. Some reviewers never send in their comments even though they agreed to perform the review so the editor has to decide whether to go with less reviews or use more time sending out the manuscript for further review. Editors need to fill their issues and the plethora of journals with overlapping interests means that many second and lower tier journals do not receive an abundance of quality manuscripts. Thus, they don't want to keep a good manuscript waiting.

4. If the journal does not have "peers" review submitted manuscripts, an editorial board with diverse backgrounds usually handles the review process. These journals normally do not have as high a reputation as the "peer review" type and articles in them are not considered as reputable or important as those in "peer reviewed" journals.

5. Once the reviews are in, the editor will virtually always send the author a letter summarizing the reviewers' comments and giving the editor's decision and recommendations.

Be ready for emotional responses to reviews!

Reviews tend to engender a disproportionately huge emotional response. Before reading the reviews, remember that this is just a manuscript - not your child. Nobody is criticizing you personally if they don't like what you wrote. Take a deep breath and have a very thick skin. You need to remain calm enough (at least after a while) and take the time to read what the reviewers said and use their comments in whatever way possible to strengthen your manuscript. Shredding or pounding them (the reviews, not the reviewers) to paste may feel good for a few minutes but then they can't be retrieved after you calm down. This deprives you of the ability to move on.

It is very rare that a manuscript is accepted without some modification. Thus, at best you can expect at least some requests for minor changes. At worst, you will get an outright rejection. Rejections are usually very carefully explained. The manuscript may be rejected because it simply does not meet the needs of the journal and the editor may suggest other journals which may be more interested in it. Usually the rejection is because the reviewers feel that there are fatal flaws in the manuscript.

A very common request from the editor is for extensive changes to be made after which the manuscript will be reconsidered. You may not feel that the reviews make any sense what-so-ever let alone being fair. The "peer review" process has a quite tarnished reputation because the editor's choices for reviewers may be your direct competitors or people who disagree with your theoretical approach and reject the paper for what you feel are inappropriate reasons. They may also be truly ignorant of what you are doing and your techniques but did the review anyway and came up with what you feel are not only factually incorrect and irrelevant but ludicrous comments. Bordens and Abbott (1991) reviewed some of the problems with peer review and quoted studies showing that a manuscript sent to 75 reviewers gets very different responses depending on the reviewer's theoretical bent and whether the data agree with the reviewer's ideas. What was written had little to do with whether the article was accepted. Other studies looked at reviews of over 150 papers reviewed by over 400 reviewers and found that inter-reviewer reliability was very low. The reviewers had such diverse comments that "it was as if they had read different papers". Other studies found interrater reliability to be about 0.25. Thus, when you feel that your paper was poorly reviewed and that the reviewers all found different problems, you may well be correct. *It is up to you to decide whether to alter and resubmit your manuscript to that journal or try someplace else.*

D. The second round: Reviewers aren't paid for doing reviews and don't get academic credit for doing them. As the reviews are normally anonymous, a few let their egos loose on defenseless authors. However, the vast majority try to do a careful review and to provide comments which they believe will result in a stronger, more readable article. Some reviewers take the time to make grammatical corrections while others restrict themselves to the scientific aspects of the manuscript. A few reviewers will identify themselves and tell the editor that it is fine if you

contact them so they can give more direct help with specific technical areas of the manuscript such as statistics or an alternate explanation for what happened in light of information you couldn't have had available to you because it hadn't been published yet.

In any case, pay very careful attention to what the reviewers say. Even if they are obviously ignorant of your techniques and the clinical problem you are treating, they are probably typical of the journal's audience. So, if they missed the idea, so will the readers.

As editors try to find subject matter experts to review your manuscript, it is common for one of the reviewers to be (or at least perceive him/her self as being) a major player in the field you are writing in. If you haven't included that person's work in the introduction, they will let you know. If you intend to resubmit to that journal, you may as well include that person's work as long as it is appropriate because you will really aggravate the reviewer if you don't and they won't be able to see past their own aggravation.

You need to make all of the changes the reviewers and editor recommend or include a very detailed justification for why you don't make a particular change. This is usually very acceptable to editors and reviewers.

If you think the review was totally incompetent, you can write a detailed letter to the editor specifying why you think this to be the case and requesting a new group of reviewers. Just saying that the reviewers had no idea what they were doing doesn't work. You have to quote from opposing sources. If the editor decides to grant you a new set of reviewers, you should expect them to be from among the absolute top experts in the technique you used and the problem you were working on. You may be even less happy with this set of reviews than the first set.

If you really feel that all of the reviews are valueless, it is probably better to send your manuscript to a different journal because editors tend to trust reviewers they know more than somebody they don't - especially if all of the reviewers thought the manuscript had multiple fatal flaws.

When you send the manuscript off to a different journal, you should strengthen it as much as possible using as many of the reviewers comments as possible. Oddly often, one or more of the reviewers chosen by the first editor are also chosen by the second editor. I get quite peeved when I receive a manuscript to review which I just reviewed for a different journal and find that none of my comments were taken seriously or defended against. Reviewers in this position, including myself, tend to let the editor know that this is the case and that pretty much ends the manuscripts consideration right there.

Some manuscripts suffer multiple rejections from journal after journal with each set of reviewers giving either similar or totally different reasons for rejecting the manuscript. When this happens, you need to be able to step back and balance feelings of perseverance with the glimmer that, just perhaps, all those reviewers are noticing a flaw whose importance you are underestimating.

When a new field is becoming established, it is frequently difficult to find a place to publish studies in the area. This leads to establishment of a specialty journal. As soon as the

journal is announced, people with multiply rejected manuscripts send them in with the virtually certain belief that they will now be accepted since those ignorant, prejudiced reviewers can't stop them now. Authors are frequently shocked when the manuscript is rejected by the new journal as well because it really does have fatal flaws the author can't seem to see.

E. Preliminary acceptance and the technical review: Once your manuscript is accepted (most reasonable ones eventually do find a home), a technical editor working for either the journal or the printer or publisher is very likely to contact you or send you a set of □proofs□. The technical editor will have gone through the manuscript with a fine tooth comb looking for problems with your citations, grammar, unclear wording, conflicts between information in tables and text, etc. You need to pay really close attention to what this person is trying to get you to do. They are doing their best to help you strengthen and clarify the manuscript but they usually aren't subject matter experts so occasionally recommend changes that don't make sense. If you simply accept the proofs as is, you may let real blunders slip through. Don't lose patience with the □nit-picking□ because the end result will be a much stronger product. In the days before most journals requested the manuscript on disk, somebody actually typed your manuscript into the journal's printing system and mistakes abounded. Things are much better now but you still need to check the proofs sentence by sentence for omissions and mistakes. It takes patience but can be well worth the trouble. I have caught several critical omissions at the last second.

A word about publication costs: Nearly all journals charge to publish color prints. Occasionally there is simply no way to translate color-coded information into greytone representations without losing crucial understanding of the technique or the results. The editor and reviewers may feel that color is required but the journal may not have the budget to pay for it. That means you may have to come up with a few thousand dollars to cover the cost. I did that once. Luckily my institution paid so it didn't come out of my pocket. I never did it again as I have always found some way around having to use color.

Some journals have □page charges□ to publish your work. You are essentially supporting the publication process. Unless such a journal is the only one which will accept your work or is such a premier journal that you just have to see your work in it so it will get the notoriety it deserves, you may as well find a free one which is just as good.

Chapter 42

Changing clinical practice based on what you did and read

You have just become aware of a new therapeutic approach and are considering trying it out on patients for whom it might be better in some way than the current approach. How can you really decide whether to give it a try? It all comes down to evaluating the credibility of the claims for the new technique's ability to do better than the techniques you are currently using.

The first thing to remember is that the vast majority of those shiny new techniques you hear about disappear or are relegated to tiny groups of fanatic supporters shortly after they make their big splash. There is rarely a pressing need to change your clinical practice unless a patient is truly suffering more harm from lack of the new technique than they are likely to get from trying it.

Some factors which point to questionable credibility - regardless of the apparent / reported strength of the evidence - include:

1. One small study with a brief follow-up.
2. Only a few studies but all done by the same group.
3. Heard about it at a meeting but no publications in peer reviewed journals.
4. The word on the street is that its terrific but no publications yet.
5. Its being pushed by someone who has something to gain by its acceptance.

Typically, this is a sales or manufacturing organization trying to get everybody to use some technique which really isn't adequately established.

6. The technique is reported to cure everybody with the problem it is aimed at.
7. The only place the "hard" data seem to be is in manufacturers sales brochures with their misleading graphs, unpublished data, and endorsements by chosen demigods. These packets also tend to contain highly selected literature which only supports use of the particular product.

Some information to look and ask for - especially when that "drug rep" is in your office and extolling the values of that miracle new product:

1. Documentation of safety studies - common side effects, rare major problems, etc. If the large studies haven't been done and you want to try the technique, ask if you can join one of the ongoing Phase III trials. If there aren't any - watch out! Nobody cares enough about the technique to actually investigate it.

2. How well demonstrated is the technique's efficacy? Has it been adequately compared with current techniques? Are the studies actually any good? Were the sample sizes adequate to tell anything at all?

3. What are the trade-offs to using the new technique such as side effects, price, convenience, etc?

4. Can you apply the new technique without significantly withholding a treatment proven to be effective in case the new one doesn't work as advertised? This is especially important when therapy has to be performed at a specific stage in the disorder's development or irreparable harm will ensue. For example, you can not delay physical therapy specific times after a stroke or contractures will occur which can't be reversed even with extensive surgery.

5. Is there a rational, believable, supported, potential mechanism through which the intervention might work? Remember that nobody really knew how aspirin effected inflammation until a few years ago. So, don't absolutely require a proven mechanism. It helps if white magic has some supporting rationale when a truly new intervention is being applied to a serious problem.

The bottom line is - how confident are you that the technique works? After reading this book, you know how to evaluate information. Establish your own set of confidence limits for the information you have, balance out what you know with what you don't and the potential for good vs. harm - then make your decision! Patient care is not the place to substitute blind acceptance for critical thinking.

If you do decide to try that new technique and are convinced that you are not likely to do any harm by applying it:

1. Try it on a few very carefully selected patients who fit the criteria for the disorder as closely as possible. This reduces the odds that you are trying the technique on a problem it wasn't meant to handle.

2. Use patients who are likely to give accurate reports of changes in their symptoms. To be absolutely blunt, this isn't the time to work with someone who needs the symptoms to maintain some psychological framework (unless the technique is aimed at curing such people of such problems).

3. Set the stage for being able to draw your own conclusions by treating your initial applications as a single group study. Take a really good baseline and do extra careful follow-ups using the most objective outcome measures you can justify and afford.

Chapter 43

Defensive reading of clinical literature

-

Does the conclusion match the raw data, the data analysis, the hypothesis, and the background?

A. Interpretation of the data -

Weaknesses in the design spread to the conclusion: While virtually all clinical fields have problems with designs due to the idiosyncratic nature of their approaches, behavioral interventions are especially prone to design weaknesses so I am picking on this field to illustrate problems to watch out for. Please note that many problems may not be avoidable but they certainly can effect the interpretation of the data. (Some of the following ideas are based on ideas from Spilker 1986)

1. Training and skills of the therapist - this effect can not be underestimated. If the therapist doesn't know how to apply the procedure, it is not likely to work. If the therapist can not train people (which includes all the parts of the patient - coach interaction) the procedure is not likely to work.

2. Intensity of the intervention - was enough training done to get an effect?

3. Was evidence for learning the technique (as opposed to change in the disorder) presented? If people don't learn, they aren't going to get better using that procedure.

4. Did the subjects continue to exert control / practice their techniques after the intervention period?

5. Were other therapeutic interventions given at about the same time as the behavioral intervention? This is the bane of biofeedback studies for such problems as low back pain. When the article is carefully analyzed, it turns out that biofeedback was just one of a half dozen interventions - both physical and behavioral - provided as a package. There is no way to realistically partial out the effects of biofeedback unless a control group was run in which everything but the biofeedback was given.

6. Was a placebo control or alternate treatment group incorporated into the procedure or was some other method for determining whether a placebo effect or time are causing the reported change? This is crucial as the placebo effect for such disorders as headache can be up to 50% and last for many months. It is frequently argued that there is no good placebo control for behavioral interventions such as biofeedback because people would rapidly realize that they were getting false feedback or no feedback. It has already been demonstrated that feeding back any relevant information - regardless of direction of training - can result in at least some symptom response apparently because simply learning to recognize what a parameter is doing leads to increased control of the parameter and, thus, perhaps, to increased control of the associated symptom. The alternatives are:

a. To give actual feedback from an unrelated physiological parameter such as theta EEG when working with rehabilitation of forearm muscle control using sEMG. Any changes are simply due to learning to sit quietly and attend to something.

b. To give actual feedback from the appropriate physiological parameter to one group and an (hopefully) ineffective, but equally likely sounding, intervention to the other group. In this design, both interventions need to be added to an accepted, moderately successful treatment or the interventions have to be given during a required waiting or baseline period prior to initiation of the standard therapy. Otherwise, the investigators may be denying appropriate care by delaying its inception.

c. To give false feedback signals from a parameter people can't sense without significant training. For example, Egner et al (2002) conducted a placebo controlled study of alpha/theta EEG training. This type of training is used to help children with ADHD. They tape recorded an entire feedback session in which a patient was able to change the amounts of alpha and theta produced and used the tape as the basis for the feedback signal. Thus the controls heard changes in the feedback signal unrelated to what their own brains' were doing. As they had no way to tell how much alpha and theta they were actually producing, they couldn't tell that they were receiving a "placebo" signal.

I frequently perform experimental interventions while subjects are on a waiting list for an accepted surgical intervention. We have occasionally had problems using a supposedly ineffective intervention as the "placebo" control because the intervention has turned out not to be as ineffective as people may have thought. Thus, double check the literature and clinical colleagues to make as sure as possible that the placebo intervention really doesn't work.

B. Evaluating the authors' interpretation of messy data: Very often the data do not seem to make much sense and the authors' explanations can seem quite far fetched. Here are some points to look for:

1. Was the correct statistical approach used? If the shotgun approach to choosing which variables to record was used (no real basis for selecting which factors might be related to the disorder) and the variables were related with each other using multiple correlations, as long as there were at least ten variables, a few should happen to correlate just by chance. The number of

articles reporting that hair color is a major predictor for the intensity of disease x is simply astounding.

2. Were sufficient subjects entered so that the results are likely to be anything but random? On the other hand, if a huge number of subjects were entered, many tests, such as parametric correlations, will find a weak, probably random or unimportant, correlation between just about any two variables.

3. Was the design appropriate to reliably detecting changes in the outcome parameter? For example, were the pre and post intervention baselines long enough with sufficient data collection points to give an accurate picture of changes in the variables?

4. Were the outcome variables actually measured by the techniques used? Was a lot of non-objective material mixed with objective data?

5. Were the subjects appropriate for the study? Were reasonable diagnostic and inclusion/exclusion criteria set and followed so the subjects were similar enough for them to have been likely to change in a similar way?

6. For blinded studies, was evidence (such as post intervention ratings) provided that the blinding worked? Was the randomization method reasonable?

7. Were the patients compliant with the intervention? Were checks built in to evaluate compliance? For example, when a physical rehabilitation study is performed which includes home practice, the devices given to the subjects need to be able to record their use. Even such simple tasks as having a patient repeatedly squeeze a ball can be checked by having a counter inside the ball.

8. Was the intervention intense enough to produce reliable reactions. A marginal dose could simply increase variability because a few subjects would respond well, more might respond just a bit, while the rest might be non-responders. Did treatment related complications confuse the outcome?

9. Who evaluated the outcomes? If several clinicians worked independently, did they train together to insure that they got the same results from the same symptoms? Did a neutral team do the evaluations or were they done by the patients' own physicians who were running the study?

10. If this study's results and conclusions are markedly different from those of the previous articles (and your clinical experience) you need to try to tease out why. If the study appears to have been reasonably well designed and powerful with reasonably well defined, appropriate subjects, you have a real dilemma. Very subtle differences in subjects and recording methodology frequently lead to enormous differences in results. The discussion section should attempt to explain why the study produced different results than its predecessors. The reality is that there is frequently no reasonable way to explain the differences. Many studies produce odd results just by chance in spite of all precautions. Others are faked. This is why repetition by independent groups is crucial.

C. The discussion/ conclusion section:

1. How are the findings related to previously cited research?
2. How are the findings related to the conceptual framework?
3. Are the generalizations appropriate or grandiose?
4. Are the conclusions valid and justified given how the study was done?
5. What are the limitations of the study?
6. What recommendations for further study are appropriate?
7. What recommendations for implementing the research are appropriate? Is more work needed before the findings can be appropriately applied to practice?
8. Are the conclusions actually based on, and justified by, the results obtained?

D. Your overall decision about the paper:

1. Do you trust what you read?
2. Would you change your clinical practice based on what you read?
3. Can you think of a way to do it better?

Chapter 44

Pitfalls in the overall research process

A. Common errors in preparing the report

(Modified and extended from (loosely based on) the "Handbook in Research & Evaluation by S. Isaac and W. Michael, 1971, Robert R. Knapp of San Diego)

1. Leave preparation of the manuscript until it is too late (e.g., too close to a presentation or abstract deadline) to do a good editing job.
2. Leave insufficient time for preparation of illustrations and statistics.
3. Literature review not related to study design or results.
4. Method section inadequate to permit replication of the study by a reader.
5. Inadequate use of illustrations and tables or illustrations inadequately labeled and tied with text.
6. Inadequate raw data presented so readers can not make their own judgements about the conclusions.
7. Conclusions not supported or marginally supported by data.
8. Wandering, useless discussion which does not tie results to clinical situation or previous literature.

B. Inaccurate estimation of the value of initial work: I can't tell you how hard it is to accurately assess an idea you had which you spent considerable time and effort (not to mention ego involvement) investigating. *It is all too easy to overstate the value of preliminary work because you believe in it and because it's new.* Unfortunately, when a new technique hits the press, just its newness gives it an air of credibility. Use your training in assessing the value of a study and of the evidence regardless of the status of the source of the information. Even your field's demigods make mistakes and occasionally back the wrong horse.

Practical exercises for Section E

1. Make a complete evaluation of two clinical articles which espouse the use of relatively untried clinical techniques for a specific clinical problem. You can use several of the articles you began evaluating for the earlier exercises if they are appropriate. Evaluate whether you should change your clinical practice based on the findings in the article.
2. Write a detailed practice protocol on a human clinical study of your choice. Include all the details you would if it was a real study that you were going to attempt. Be sure to include the statistics and budget sections as well as the consent form.

Section F

Glossary of research and statistical terms

This glossary is modified from one prepared between 1994 and 1997 by the Department of Clinical Investigation at Brooke Army Medical Center located at Fort Sam Houston in San Antonio, Texas. I appreciate their permitting the use of their glossary.

accessible population the population of subjects available for a particular study; often a nonrandom subset of the target population.

accidental sampling selection of the most readily available persons as subjects in a study; also known as convenience sampling.

accuracy validity; the proportion of all test results both positive and negative that are correct.

alpha (α); the probability of making a Type I error.

alternative hypothesis a statistical hypothesis that disagrees with the tested (null) hypothesis.

analysis of covariance (ANCOVA) a statistical procedure used to test the effect of one or more treatments on different groups while controlling for one or more extraneous variables (covariates).

analysis of variance (ANOVA) a statistical procedure for testing the effect of one or more treatments on different groups by comparing the variability between groups to the variability with groups.

anonymity protection of the participants in a study such that even the researcher cannot link them with the information provided.

assumptions basic principles that are accepted as being true on the basis of logic or reason, without proof or verification.

attribute variables preexisting characteristics of the entity under investigation, which the researcher measures (e.g., gender, age, level of education).

beta (β); the probability of making a Type II error.

bias any influence that produces a distortion in the results of a study; systematic error.

bivariate statistics statistics derived from the analysis of two variables simultaneously for the purpose of assessing the relationship between them.

case study in-depth analysis of a person, group, institution, or other social unit; an empirical inquiry that investigates a contemporary phenomenon within its real-life context, when the boundaries between phenomenon and context are not clearly evident, and in which multiple sources of evidence are used.

case control study a study which starts with an outcome (e.g. disease), selection of a sample from a population of patients with the outcome (cases) and another sample from a population without the outcome (controls), and comparison of the levels of the predictor variables in the two samples to determine which ones are associated with the outcome; retrospective studies.

causal relationship a relationship between two variables such that the presence or absence of one variable (the "cause") determines the presence or absence, or value, of the other (the "effect").

cell the intersection of a row and column in a table with two or more dimensions.

central tendency a statistical index of the "typicalness" of a set of scores that comes from the center of the distribution of scores. The three most common indices of central tendency are the mode, the median, and the mean.

chi-square test a non-parametric test of statistical significance used to assess whether or not a relationship exists between two nominal-level variables. Symbolized by X^2 .

clinical trials a type of cohort study in which the conditions of study (selection of treatment groups, nature of interventions, management during follow-up, etc.) are specified by the investigator for the purpose of making unbiased comparisons; intervention or experimental studies.

closed-ended question a question that offers respondents a set of mutually exclusive and jointly exhaustive alternative replies, from which the one that most closely approximates the "right" answer must be chosen.

cluster sampling a form of multistage sampling in which large groupings (clusters) are selected first, with successive subsampling of smaller units.

codebook the documentation used in data processing that indicates the location and values of all the variables in the data file.

coding the process of transforming raw data into standardized form for data processing and analysis.

coefficient alpha (Cronbach's alpha) a reliability index that estimates the internal consistency or homogeneity of a measure composed of several items or subparts.

cohort a defined group of study subjects, often similar in age. The usual idea is that the cohort consists of everyone who had some procedure so the entire group, rather than some distorted sample is followed throughout the life history of the study.

cohort study a type of trend study that focuses on a specific subpopulation (e.g. an age related group) from which different samples are selected at different points in time. A group of people (a cohort) is assembled, none of whom has experienced the outcome of interest; subjects are classified according to those characteristics that might be related to outcome; then observed over time to see which of them experience the outcome. Also called longitudinal, prospective, incidence studies.

comparison group a group of subjects whose scores on a dependent variable are used as a basis for evaluating the scores of the target group or group of primary interest. The term "comparison group" is generally used instead of "control group" when the investigation does not use a true experimental design.

concurrent validity the degree to which an instrument can distinguish individuals who differ on some other criterion measured or observed at the same time.

confidence interval the range of values within which the parameter has a specified probability of lying; the interval around an effect size (usually a 95% confidence interval) that is interpreted as: if the study is unbiased, there is a 95% chance that the interval includes the true effect size. The narrower the confidence interval, the more certain one can be about the size of the true effect. An alternative way of expressing statistical significance: if the value corresponding to no effect falls outside 95% confidence interval, the results are statistically significant at the .05 level.

confidentiality protection of participants in a study such that their individual identities will not be linked to the information they provide and divulged in public.

construct validity the degree to which an instrument measures the concept under investigation

content analysis a procedure for analyzing written, verbal, or visual materials in a systematic and objective fashion, typically with the goal of quantitatively measuring variables.

content validity the degree to which the items in an instrument represents content appropriate to the behavior under study.

contingency tables the manner for displaying chi-square data which gives the observed frequencies of individuals falling into each category.

control procedures used to hold constant possible confounding influences on the dependent variable.

control group the group in an experimental investigation that does not receive the intervention; the comparison group.

convenience sampling selection of the most readily available persons or units as subjects in a study; also known as accidental sampling.

correlation the degree of association between two variables; that is, a tendency for variation in one variable to be related to variation in another variable.

correlation coefficient an index that summarizes the degree of relationship between two variables. Correlation coefficients typically range from +1.00 (for a perfect direct or positive correlation) to 0.0 (no relationship) to -1.00 (for a perfect inverse or negative relationship).

counterbalanced designs designs in which experimental control is achieved by entering all subjects into all treatment conditions in systematic order so that carry-over effects can be ascertained; one type of design is the Latin Square design.

covariate a variable that is statistically controlled in analysis of covariance; an extraneous, confounding influence on the dependent variable.

correlational research nonexperimental investigations that explore the interrelationships among variables without any intervention on the part of the researcher.

criterion validity the degree to which scores on an instrument are correlated with some external criterion. Criterion validity is established by showing that the measurement predicts a directly observable phenomenon.

critical region that portion of the area under the curve which includes those values of a statistic that lead to rejection of the null hypothesis.

cross-sectional studies a study based on observations of different age or developmental groups at a single point in time for the purpose of inferring trends over time.

crosstabulation a determination of the number of cases occurring when simultaneous consideration is given to the values of two or more variables (e.g., sex-male/female crosstabulated with smoking status - smoker/non-smoker). The results are typically presented in a table with rows and columns divided according to the values of the variables.

data the values of variables measured in the course of a study (singular is datum).

degrees of freedom a concept used in tests of statistical significance, referring to the number of sample values that cannot be calculated from knowledge of other values and a calculated statistic (e.g. by knowing a sample mean, all but one value would be free to vary); degrees of freedom (DF) is usually $N-1$, but different formulas are relevant for different tests.

Delphi technique a data collection technique developed by the Rand Corporation for obtaining judgements from a panel of experts. The experts are questioned individually, a summary of the judgements is circulated to the entire panel, the experts are questioned again, with further iterations until consensus is reached.

delta the magnitude and direction of the difference between populations tested.

dependent variable the outcome variable of interest; the variable that is hypothesized to depend on or be caused by another variable (called the independent variable).

descriptive research studies designed to describe the characteristics of persons, situations, or groups, and the frequency with which certain phenomenon occur.

descriptive statistics statistics used to describe and summarize the data set from the sample; usually consists of measures of central tendency (mean, median, mode), measures of dispersion or variability (range, standard deviation), and the presentation of this information in the form of graphs and tables; may include correlation results.

dichotomous variable a variable having only two values or categories (e.g., gender).

directional hypothesis a hypothesis that makes a specific prediction about the direction (i.e., positive or negative) of the relationship between two variables.

discriminant analysis a statistical procedure used to predict group membership or status on a categorical (nominal level) variable on the basis of two or more independent variables.

double-blind a condition in which allocation of the treatment in experimental studies is unknown to both the experimenter and subject until completion of the study. Observational studies may also be blinded in various ways; for example, interviewers in a case-control study may not be told whether subjects are cases or controls.

effect size the magnitude of the "real" effect to be detected.

epidemiology the field of science or research discipline that deals with the frequency, distribution, and determinants of disease in a population.

error of measurement the degree of deviation between true scores and obtained scores when measuring a characteristic; also called error variance.

evaluation research research that investigates how well a program, practice or policy is working.

experimental mortality a threat to internal validity in which there is a differential loss of subjects from the comparison group; especially a problem in studies carried out over a long period of time.

ex post facto research research conducted after the variations in the independent variable have occurred in the natural course of events; a form of nonexperimental research in which 'causal' explanations are inferred "after the fact".

experiment a research study in which the investigator manipulates an independent variable (the treatment) to test the effect on one or more dependent variables, administers the treatment to some but not to others (control group), controlling for extraneous influencing factors by randomly assigning subjects to the control and experimental groups; designs that meet these three criteria (manipulation, control group, randomization) are usually identified as true experimental designs to differentiate them from experimental designs in which one or more of the criteria are missing.

experimental group the group in an experimental design that receives the treatment or intervention being tested.

exploratory research an extension of descriptive research that focuses on the discovery of relationships and pursues questions such as: what factor(s) influence, affect, cause, or relate to a phenomenon of interest.

external validity the degree to which the results of a study hold true for settings or samples other than the ones studied; the generalizability of the results.

extraneous variables variables that confound the relationship between the independent and dependent variable(s) and that need to be controlled either in the research design or through statistical procedures.

factor analysis a statistical procedure for reducing a large set of variables into smaller sets of variables with common characteristics or underlying dimensions.

factorial design an experimental design in which two or more independent variables are simultaneously manipulated; this design permits an analysis of the main effects of the independent variables separately, plus the interaction effects of these variables.

field study a study in which the data are collected "in the field" from persons in their normal roles, rather than as subjects in a "laboratory" study.

Fisher's Exact test an exact calculation of alpha used to determine if two groups are independent with respect to a nominal independent variable. It replaces the chi-square test when the sample size is less than 20 or the expected frequency in any cell of the contingency table is less than 5.

follow-up study a study undertaken to determine the subsequent development of persons with a specified condition or who have received a specified treatment.

Friedman test a nonparametric test of the difference in the ranks of scores for three or more related sets of scores (i.e., dependent variable is measured at the ordinal level). A non-

parametric ANOVA.

frequency distribution a systematic presentation of numerical values from the lowest to the highest, together with a count of the number of times each value was obtained.

generalizability the degree to which the research procedures justify the inference that the findings represent something beyond the specific data upon which they are based; i.e. the inference that the findings can be generalized from the sample to the entire population.

Halo effect the tendency of a rater to be influenced by one or more characteristics of the subject in rating other nonrelated characteristics; bias on the part of the rater which can be identified by conducting **interrater reliability** tests.

Hawthorne effect the effect on the dependent variable caused by subjects awareness that they are "special" participants under study.

heterogeneity the degree to which subjects or objects in a study are dissimilar with respect to some attribute; high variability.

historical research the systematic collection and critical evaluation of data relating to past events for the purpose of shedding light on present behaviors or practices.

history a threat to the internal validity of a study; refers to the occurrence of events between the first and second measurement periods, and external to the treatment but concurrent with it, which can affect the outcome (dependent) variable.

homogeneity (1) in terms of the reliability of an instrument, the degree to which the subparts are internally consistent (i.e. are measuring the same critical attribute); (2) generally, the degree to which subjects or objects are similar (i.e. characterized by low variability).

hypothesis a statement of predicted relationships between variables under investigation; hypotheses lead to empirical studies that seek to confirm or disconfirm those predictions.

incidence the fraction or proportion of a group initially free of the condition that develops it over a given period of time. **Incidence studies** identify the population free of the event of interest and then follow them through time with periodic examinations to determine occurrences of the event; cohort studies.

independent variable the variable that is believed to cause or influence the dependent variable; in experimental research, the variable that is manipulated.

inferential statistics statistics that permit us to infer whether relationships observed in the sample are likely to occur in the population at large; a quantitative science that, based on assumptions about the mathematical properties of the data, allows calculations of the probability that the results could have occurred by chance alone.

informed consent the ethical requirement that researchers obtain the voluntary participation of subjects after informing them of possible risks and benefits of participating in a study.

instrument refers to equipment used to collect the data in a study. For example, biomedical instruments are used to collect physical and physiological data (specific gravity, temperature); paper and pencil instruments such as tests, questionnaires, interview schedules are used to collect psychological, cognitive, social data (attitudes, perceptions, knowledge, life history reports, etc).

instrumentation effects changes or deterioration in the measuring instrument (such as biomedical equipment) with repeated use that can produce changes in the data collected resulting in errors of measurement; can occur with observers or scorers who become fatigued over time; these factors need to be managed through periodic calibration of equipment, rest periods for observers, etc. in order to reduce measurement error.

interaction effect the effect on a dependent variable of two or more independent variables acting in combination rather than as unconnected factors.

internal consistency a form of reliability, referring to the degree to which the subparts of an instrument (e.g. questionnaire) are all measuring the same attribute or dimension.

internal validity the degree to which it can be inferred that the experimental treatment (the independent variable) rather than uncontrolled, extraneous factors, are responsible for the resulting effect on the outcome (dependent) variable. Internal validity is related to the conclusions about the findings (e.g. causal vs relational) based on the design and the extent to which bias or confounding (extraneous) factors are controlled.

interrater reliability the degree to which two raters, operating independently, assign the same ratings for an attribute being measured.

interval measure a level of measurement in which an attribute of a variable is rank ordered on a scale that has equal distances between points on that scale but in which there is an arbitrary zero (e.g. Fahrenheit scale). Interval scales are continuous (can take on any value in a continuum) or discrete (can take on only specific values).

intervention in experimental research, the experimental treatment or manipulation that is being tested.

interview a method of data collection in which one person (the interviewer) asks questions of another person (a respondent) using an interview schedule; interviews are usually conducted either face-to-face or by telephone; face-to-face interviews are usually tape recorded (with the respondents permission) in order to facilitate verification of responses by interrater agreement, and to facilitate coding, and content analysis.

item a term used to refer to a single question on a test or questionnaire, or a single statement on a scale.

judgmental sampling a type of nonprobability sampling method in which the researcher selects subjects for the study on the basis of personal judgment about which ones will be most representative or productive; also referred to as purposive sampling; the type of sampling that may result from specifying limited inclusion or exclusion criteria for subjects to participate in the study.

key informant a person well-versed in the phenomenon of research interest and who is willing to share the information and insight with the researcher.

known-groups technique a technique of estimating the construct validity of an instrument through an analysis of the degree to which the instrument separates groups that are predicted to differ on the basis of some theory or known characteristic from those who do not have the characteristic.

Kruskal-Wallis test a nonparametric one-way analysis of variance to test the difference in the rank scores of three or more independent groups (i.e., dependent variable measured at the ordinal level).

kurtosis the measure of relative peakedness or flatness of a distribution curve; on an SPSS printout, 0 indicates a normal curve, a positive number indicates a narrow, peaked curve (fewer than normal proportion of cases fall in the tails), and a negative number indicates a flatter curve (larger proportion of cases fall in the tails).

Latin Square (see counterbalanced designs)

level of significance (significance level, alpha) the likelihood that results obtained in an analysis of sample data were caused by chance at a specified level of probability (usually $p = .01$ or $.05$); the probability of a given result.

likelihood ratio for a particular value of a diagnostic test, the probability of that test result in the presence of disease divided by the probability of the result in people without the disease.

Likert scale a type of composite measure of attitudes that involves the summation of scores on a set of items (statements) to which respondents are asked to indicate their degree of agreement or disagreement. Originally five categories of agreement-disagreement response alternatives were used; however seven, six, and ten point scales have also been constructed with differences of opinion concerning their appropriateness.

linear regression (simple) a method of analysis which uses one variable to predict a second variable.

literature review a critical summary of research on a topic of interest, prepared to put a research problem in context, identify gaps and weaknesses in prior studies, and to justify a new investigation.

log-linear analysis non-parametric statistical analysis procedures used to study the relationship

among variables that are nominal; does not distinguish between independent and dependent variables; used in the case of multiple categorical variables and estimates effects and interaction similar to factorial analysis of variance.

logistical regression non-parametric multiple regression procedures used to predict the probability of a single binary dependent variable from a number of independent variables.

longitudinal study a study designed to collect data at more than one point in time, in contrast to a cross-sectional study.

manipulation an intervention or treatment introduced by the researcher in an experimental or quasi-experimental study.

Mann-Whitney U-test a nonparametric statistical test employed as an alternative to the t-test when dependent variable measurements are at the ordinal level; tests the difference between ranks of scores in two independent groups.

Mantel-Haenzel Chi square a nonparametric test in which the dependent variable involves frequencies in categories and there are two categorical independent variables.

matching the pairing of subjects on one group with those in another group based on their similarity on one or more dimensions, done in order to enhance the overall comparability of groups; when matching is performed on the context of an experiment, the procedure results in a randomized block design.

maturation a threat to internal validity of a study, especially those involving data collection over extended periods of time, in which processes within the subjects operating as a function of time could influence the response on the dependent variable (e.g., growing older, hungrier, more tired, etc.).

McNemar test a non-parametric test for comparing two dependent groups where the dependent measure is at the nominal or ordinal level of measurement and there are two measures on each subject (e.g. pre-, post-), or other situations in which individual measurement in one sample can be paired with a particular measurement in the other; comparable to the paired t-test.

mean a descriptive statistic that is a measure of central tendency, computed by summing all scores and dividing by the number of subjects; the average.

measurement the assignment of numbers to objects according to specified rules to characterize quantities of some attribute.

median a descriptive statistic that is a measure of central tendency, representing the exact middle score or value in a distribution of scores; the value (the 50th percentile) above and below which 50% of the scores lie.

mode a descriptive statistic that is a measure of central tendency and which indicates the value

that occurs most frequently in a distribution of scores.

model a symbolic representation of concepts or variables, and interrelations among variables.

multiple regression the statistical techniques for using several independent variables to predict a dependent variable.

multistage sampling a sampling strategy that proceeds through a set of stages from larger to smaller sampling units (e.g. from states, to hospitals, to administrators).

Multitrait-multitrait method method of establishing the construct validity of an instrument **matrix approach** that involves the use of multiple measures for a set of subjects. The target instrument is valid to the extent that there is a strong relationship between it and other measures purporting to measure the same attribute (convergence) and a weak relationship between it and other measures purporting to measure a different attribute (discriminability).

multivariate analysis statistical procedures used to analyze the relationship among three or more variables, where there is usually more than one dependent variable.

N used to designate the total number of subjects in a study.

n used to designate the number of subjects in a subgroup or a cell of a study.

needs assessment a study in which a researcher collects data for estimating the needs of a group, community, or organization; usually used as a guide to resource allocation.

negative predictive value the probability that a person who is identified by a diagnostic test as healthy really is healthy; that is, the probability of a "true negative". See also positive predictive validity, sensitivity, specificity.

negative relationship a relationship between two variables in which there is a tendency for higher values on one variable to be associated with lower values on the other (e.g. as the temperature increases, people's productivity decreases); also referred to as an inverse relationship.

nominal measure the lowest level of measurement which involves the assignment of characteristics into categories (e.g. males, category 1, females, category 2) without any inherent order; also referred to as categorical data.

nondirectional hypothesis a research hypothesis that does not specify the direction (i.e. positive or negative) of the relationship between variables.

nonequivalent control group a comparison group that was not developed on the basis of random assignment of subjects to the groups; used when randomization is not feasible; there is no way of assuring the initial equivalence among different groups unless data is collected before the treatment is administered and any differences between groups can be determined statistically.

nonexperimental research studies in which the researcher collects data without introducing any new treatments or changes.

nonparametric statistics a general class of inferential statistics that does not involve rigorous assumptions about the distribution of the critical variables; used when samples are small and when the data are measured on nominal or ordinal scales.

nonprobability sampling the selection of subjects or sampling units from a population using nonrandom procedures (e.g., convenience, quota, purposive sampling).

normal distribution a theoretical distribution that is bell-shaped and symmetrical in which 68% of the area under the curve lies within 1 standard deviation (SD) above or below the mean; 95% of the area lies within 2 SD; and 99% of the area lies within 3 SDs.

null hypothesis the statistical hypothesis that states there is no relationship between the variables under study; sometimes used in place the alternative hypothesis in tests of statistical significance; the hypothesis to be rejected in order to accept the alternate hypothesis.

observational research (1) studies in which the data are collected by means of observing and recording values without controlling or manipulating an independent variable; the researcher gathers data by observing events as they happen without taking an active part in what takes place; or (2) studies in which data are collected by direct observation through the senses recording behaviors or activities of interest, such as physiologic conditions, verbal and non-verbal behavior, skill & performance, environmental characteristics, using real time or time-sampling methods to capture the total behavior or a sampling of the behavior.

odds the ratio of 2 probabilities; $\text{odds} = \frac{\text{probability of an event}}{1 - \text{probability of the event}}$.

odds ratio a measure of risk, usually obtained from case-control studies and (when studying rare diseases) mathematically close to relative risk.

open-ended question a question in an interview or questionnaire that does not restrict the subject's answers to preestablished alternatives.

operational definition the definition of a concept or variable in terms of the operations or procedures by which it is to be measured.

ordinal measure a level of measurement that produces rank orders of a variable from low to high or high to low along some dimension; the magnitude of differences between ranks is not consistent.

outcome measure the dependent variable, the variable the researcher is attempting to predict.

outliers in data analysis, the cases with large residuals.

p-value a statistical estimate of the probability that a finding is due to chance; a finding of a p-value less than five percent (.05) or less than one percent (.01), is, by convention, called "statistically significant"; the same as the alpha level or level of significance.

paired t-test parametric statistical test employed to test the difference between the means of two related (e.g. matched) groups or sets of scores when the dependent variable is at least at the interval level.

panel study a type of longitudinal study in which the same subjects are used to provide data at two or more points in time.

parameter a characteristic of a population (e.g. the mean age of all US citizens).

parametric statistics a class of inferential statistics that involves assumptions about the distribution of the variables, the estimation of a parameter, and the use of interval or ratio measures.

Pearson product moment correlation a parametric test of the correlation between two variables measured at the interval or ratio level.

pilot study a small-scale version, or trial run, done in preparation for a major study.

placebo an intervention that is intended to be indistinguishable from the active treatment - in physical appearance, color, taste, odor, etc - but does not have a specific, known mechanism of action.

population the entire set of people (or objects) having some common characteristic(s).

positive relationship a relationship between two variables in which there is a tendency for high values on one variable to be associated with high values on the other.

positive predictive value the probability that a person who is identified by a diagnostic test as diseased, really is diseased; that is, the probability of a "true positive". See also negative predictive validity, sensitivity, specificity.

power the probability of rejecting the tested (null) hypothesis when it is false, that is, when an alternative hypothesis is true; denoted by $1 - \beta$, where β is the probability of a type II error.

power analysis determining the statistical power of various configurations of a statistical test, alpha level, sample size, and effect size for the purpose of (1) enhancing the effect size to make the design as efficient as possible, (2) determining the necessary sample size, and (3) relaxing the error risk criteria (alpha and beta) if necessary to accommodate limits on sample size or effect size.

predictive validity the degree to which an instrument can predict some criterion measure obtained at some future time; the probability of disease given the results of a test.

pre-experimental design a research design that does not include controls to compensate for the absence of either randomization or a control group.

pretest the collection of data prior to the experimental intervention; sometimes referred to as baseline data.

prevalence studies one-shot examinations or surveys of a population of individuals including cases and noncases; cross-sectional studies.

probability a theory concerned with the probable outcome of experiments; the number of favorable cases divided by the total number of (equally possible) cases or $p = f/(f+u)$, where p is probability, f is the number of favorable cases, and u is the number of unfavorable cases; $p = \frac{\text{area under a portion of a curve}}{\text{total area under the curve}}$; used to total area under the curve express sensitivity, specificity, and predictive values - the proportion of people in whom a particular characteristic is present.

probability sampling the selection of subjects or sampling units from a population using random procedures; examples include random sampling, cluster sampling, and systematic sampling.

probing usually in the use of interviews, eliciting more useful or detailed information from a subject than was volunteered during the first reply.

proposal a document specifying what the researcher proposes to study including the research problem, its importance, procedures planned for solving the problem, and, when funding is sought, how much the research will cost.

prospective studies a study that begins with an examination of presumed causes (e.g. cigarette smoking) and then goes forward in time to observe presumed effects (e.g. lung cancer).

purposive sampling a type of nonprobability sampling method in which the researcher selects subjects for the study on the basis of personal judgement about which ones will be most representative or productive; also referred to as judgmental sampling.

Q-sort a method of scaling in which the subject sorts statements into a number of piles (usually 9 or 11) according to some bipolar dimension (e.g. most like me/least like me; most useful/least useful).

qualitative analysis the nonnumerical organization and interpretation of data for the purpose of discovering important underlying dimensions and patterns of relationships.

quantitative analysis the manipulation of numerical data through statistical procedures for the purpose of describing phenomena or assessing the magnitude and reliability of relationships among them.

quasi-experimental a study in which subjects cannot be randomly assigned to treatment

conditions but in which there is manipulation of an independent variable and certain controls are used to increase the internal validity of the results.

questionnaire a method of gathering self-report information from subjects through self-administration of questions in a paper-and-pencil format.

quota sampling the nonrandom selection of subjects in which the researcher prespecifies characteristics of the sample to increase its representativeness.

random-number table a table of 4000 random numbers generated at the Rand Corporation, in the range of 0 to 100, and organized in 40 sets of 100 each; portions of this table or the generation of tables by random number computer programs are used in random sampling.

randomization the assignment of subjects to treatment conditions in a such a manner that each person in a defined population has an equal chance of being assigned to any of the groups; accomplished by using a table of random numbers or a computer generated list of random numbers.

random selection a manner of selecting people for a study, in which each person in a defined population has an equal chance of being chosen; accomplished through use of a table of random numbers or a computer generated list of random numbers.

random variation the divergence of an observation on a sample from the true population value, due to chance alone.

range a measure of variability consisting of the difference between the highest and lowest values in a distribution of scores.

ratio measure level of measurement in which there are equal distances between score units and there is a true meaningful zero point (e.g. age).

relative risk the probability of some untoward event; the likelihood that people who are without a disease, but exposed to certain factors ("risk factors") will acquire the disease.

reliability the degree of consistency or dependability with which an instrument measures the attribute it was designed to measure; the probability that a test or measure will produce the same results on repeated measurement; other terms sometimes used to denote reliability are: sensitivity, precision, reproducibility.

representative the extent to which the key characteristics of the sample approximate those of the population.

replication the duplication of research procedures in a second investigation for the purpose of determining if earlier results can be repeated.

research systematic inquiry that uses orderly scientific methods to answer

questions or solve problems.

research design the overall plan for collecting and analyzing data, including specifications for enhancing the internal and external validity of the study.

residual the difference between the observed and predicted values of the dependent variable, or what is left over after the model is fitted to the data.

response rate the rate of participation in a survey; calculated by dividing the number of persons participating by the number of persons sampled.

response bias the measurement error introduced by the tendency of some people to respond to items in characteristic ways (e.g., always agreeing) independently of the item's content.

retrospective study a study that begins with the manifestation of the dependent variable in the present (e.g., lung cancer) and then links this effect to some presumed cause occurring in the past (e.g. cigarette smoking).

risk the probability of some untoward event; the likelihood that people who are without a disease, but exposed to certain factors ("risk factors") will acquire the disease.

risk ratio the ratio of incidence in exposed persons to incidence in nonexposed persons; relative risk.

sample a subset of a population selected to participate in a research study.

sampling the process of selecting a portion of the population to represent the entire population.

sampling bias distortions that arise from the selection of a sample that is not representative of the population from which it was drawn.

sampling frame a list of all the elements in the population, from which the sample is drawn.

scale a composite measure of an attribute, consisting of several items that have a logical or empirical relationship to each other; involves the assignment of a score to place subjects on a continuum with respect to the attribute.

Scheffe' test a multiple comparison method to determine where the difference lies when a statically significant difference is found among more than two groups.

selection bias a threat to the internal validity of a study that results from pre-treatment differences between experimental and comparison groups.

self-report any procedure for collecting data that involves a direct report of information by the person who is being studied (e.g., by interview or questionnaire).

semantic differential a technique used to measure attitudes that asks respondents to rate a

concept of interest on a series of seven-point (but un-numbered) bipolar rating scales; adjective pairs cluster along dimensions of evaluation, potency, and activity (e.g. good/bad, strong/weak, fast/slow).

sensitivity the probability that a person with a given disease will be correctly classified by some diagnostic test

Solomon four group design a true experimental design in which subjects are randomly assigned to one of four groups, two groups are pre- and post-tested, one of which gets the experimental treatment; and the other two groups are only post-tested, one of which gets the experimental treatment.

specificity the probability that a person without a given disease will be correctly classified by some diagnostic test.

significance level (alpha) the probability that an observed relationship could be caused by chance (i.e. because of sampling error); the probability of making a type I error (rejecting the null hypothesis when it is true); significance at the .05 level indicates the probability that a relationship of the observed magnitude would be found by chance only 5 times out of 100; the lower the significance level, the lower the likelihood of making a type I error.

skewness a distribution is skewed if it is not symmetrical but has more cases on one side of the tail than the other; this disproportion causes a "tail" to be formed at one end of the distribution; if the tail extends to the right, the distribution is positively skewed; a negative skew has the tail extended to the left; in SPSS programs: 0 indicates a normal distribution, a positive value indicates a positively skewed distribution and a negative value indicates a negatively skewed distribution.

split-half technique a method for estimating the internal consistency (reliability) of an instrument (usually a questionnaire or test) by correlating scores on half of the measure with scores on the other half.

standard deviation (SD, s , or σ) a measure of variability; the square root of the variance; the root mean square of the difference between all of the measures in the set of data and the mean value for the set.

standard error the standard deviation of a sampling distribution.

statistic an estimate of a parameter (about a population) calculated from sample data.

statistical regression a threat to internal validity in which there is a tendency for subjects scoring at the extremes of a distribution to exhibit less extreme scores on retesting.

statistical significance a term indicating that the results obtained in an analysis of sample data are not likely to have been caused by chance, at some specified level of probability (e.g., .05, .01).

strata subdivisions of the population according to some characteristic.

stratification a general strategy for dealing with a variable that may be a confounder, in which subjects are separated into groups according to their level of the confounding variable, and analyzed within those groups.

stratified random sampling the separate random selection of subjects from two or more groups (strata) of subjects from the same population.

subject a person who participates and provides data in a study.

survey research a type of nonexperimental research that focuses on obtaining information about the status quo of some situation, often through direct questioning of individuals in a sample.

survival analysis (life tables); procedures to determine the likelihood, on the average, that patients with a given condition will experience an outcome at any point in time.

systematic sampling the selection of subjects such that every k th (e.g. every 10th) person or element in a sampling frame or list is chosen.

target population the entire population in which the researcher is interested and to which he or she would like to generalize the results of the study.

test-retest reliability the stability of an instrument determined by correlating the scores obtained on repeated administrations; an indication of the instruments susceptibility to influencing factors (and therefore, lack of stability) from one administration to the next.

test statistic the quantity x on which the decision to accept or reject the null hypothesis is based.

theory an abstract generalization that presents an explanation about the relationships among phenomena (variables, concepts).

time sampling in observational research, the selection of time periods during which observations will take place.

time-series design a quasi-experimental design that involves the collection of information over an extended period of time, with multiple data collection points both prior to and after the introduction of a treatment.

treatment the experimental manipulation of an intervention.

trend study a form of longitudinal study in which different samples from a population are studied over time with respect to some phenomenon.

true score a hypothetical score that would be obtained if a measure were infallible; the portion of the observed score not due to random or measurement error.

t-test a parametric statistical test used for analyzing the difference between two means; Student's t.

two-tailed test a test on the hypothesis that the critical region of value extremes are in both directions.

Type I error a decision to reject the null hypothesis when it is true (i.e., the conclusion that a relationship exists when in fact it does not).

Type II error a decision to accept the null hypothesis when it is false (the researcher concludes that no relationship exists when in fact it does).

validity the degree to which an instrument measures what it was intended to measure; the extent to which the results of a study correspond to what is actually happening; accuracy.

variability the degree to which values on a set of scores are widely different or dispersed.

variable a characteristic or attribute of a person or object that varies within the population under study (e.g., age, heart rate).

variance a measure of variability or dispersion; the square of the standard deviation.

Section G

Typical protocol format and structure

(Note: this protocol is not in the format for any particular institution - be sure to get the format your institution is currently using. Remember that formats can change radically from year to year so don't use an old protocol as the basis for your current effort.)

Human Use Protocol (*center at top of page*)
to be conducted at _____

1. Date submitted to the Human Use Committee:
2. Title: *Must be descriptive of the work attempted*
3. Investigators: *All roles must be defined in detail including % time spent on protocol. Further definition of roles can be in the method section.*
 - a. Principal Investigator:
 - b. Associate Investigator(s) and consultants: *If you do not have the expertise to perform some aspect of the protocol (such as data analysis), you need to work with people who do. Failure to show appropriate expertise being available, as demonstrated by the investigators resumes, will result in the protocol being returned without consideration.*
 - c. Staff Mentor (*if appropriate*):
 - d. Technical Staff:
4. Summary: *This is the same as the summary found in an article, It must include the hypothesis, identify the population studied, the methods, results (including basic statistics) and conclusion. This section is normally limited to 500 words.*

5. Facilities to be used: *If facilities outside your institution are to be used, a letter of agreement from the director of that facility must be attached. If the other facility requires a separated human use or research use approval from that provided by this committee, you must inform this committee here. You can not begin your research until the other organization has approved the study if they need to do so. This committee's approval does not come into force and no letter of approval will be issued until all outside approvals are provided to the committee.*

6. Time course of study:

a. Anticipated start date:

b. Anticipated completion date:

7. Hypothesis (es): *this (ese) is (are) stated as answerable questions. Each hypothesis is repeated in the statistical analysis section with a detailed explanation of how it will be tested.*

8. Objective(s): *these are operationally defined goals and how they will be reached. Some organizations literally require step by step definitions of how you will achieve each sub-goal in the study such as recruiting subjects, etc.*

9. Medical application / significance to the field: *The committee is less likely to approve the use of time and resources for projects which do not appear likely to advance the field*

10. Status: *This is similar to the introduction to an article but is in far greater depth. You must demonstrate that you understand the status of work associated with your proposal and how your project fits into and advances the field. The background for all psychometric tools, equipment, etc. you intend to use must be fully explained. For example, if you are going to have subjects use a ten number visual analog scale to rate nausea, you must review the literature supporting the use of this tool and show that it is optimal for your study and population. Interdigitate citations to work quoted by placing the first author's last name and the year of the publication in parenthesis. Do not number the references. <This section is limited to 10 pages.>*

11. Study Plan: *This is a detailed description of what you want to do. Someone not involved with planing your project should be able to duplicate your work exactly from this description. If appropriate, a time line for reaching intermediate goals is provided. <This section is limited to 15 pages.>*

a. Subjects:

(1) Number of subjects: *Either justify the number of subjects based on a power analysis , design the study so it begins with a pilot which would permit you to determine the eventual number of subjects required, or justify why this can't be done - as in cases where an entire population will be studied (e.g. every human in NY who smokes for a population study).*

(2) Age range of subjects: *Explain your choice.*

(3) Sex and racial composition of subjects: *It is now illegal to exclude any group or sex without a specific reason (e.g. you do not have to include anyone other than black females if you are studying availability of care for stress related urinary incontinence among black*

females).

(4) Source and availability of subjects: *You need to provide evidence that you can actually get sufficient subjects meeting your study criteria. This evidence can take the form of the results of a medical records audit from your clinic, etc.*

(5) Inclusion, exclusion, and diagnostic criteria: *This is a crucially important section. You must define exactly who can and can not participate in your study and you must specify why this is so. Saying that people with psychiatric problems can not participate is not sufficient. You must justify why they can't. You must be absolutely explicit about how you intend to define who can participate if you are working with a disorder. For example, if you were working with people having classic migraine headaches, you would reference some recognized source such as the ad hoc committee for definition of headaches and you would state their criteria for inclusion as you intend to apply them. For example, a patient must have headaches at least x times per month but not more than y times per week, the patient must have some type of aura prior to initiation of the headache, they must vomit during x out of y headaches (and the vomiting must relieve the pain), etc.*

(6) Identification of subjects: *How you will record the subjects' identities and protect their privacies.*

b. Evaluations before entry: *How you plan to test the subjects to determine whether they are eligible for your study. This includes medical examinations (be specific - do not say a medical screening will be performed), psychometric evaluations (state specific tests and cut off criteria), medical records evaluations, etc.*

c. Methods: *State exactly and precisely what you are going to do, how you will do it and the order in which it will be done! A diagram showing the overall structure of the study and an individual subject's participation and flow through the study is usually very helpful. Specify and justify every technique used. For example, if you are recording low back muscle tension, say how often you will do it, why the sessions are spaced as you are spacing them, exactly how you will do the recordings (method of sensor application, impedance cut offs, bandwidths, subject positions, type of data recorded, etc.), etc. You must justify your choices either here or in the introduction. For example, you must justify your choice of using bandwidth y to record from muscle x . You would quote your source for stating that the power-spectrum for muscle x falls within bandwidth y while it is in the conditions under which you are recording (static with no opposed force, etc.). If you are randomizing subjects into groups, specify how they will be randomized. If you are using medications, be very specific about who is keeping track of the subject's medical conditions, how you know they are taking the drugs, etc. For project Medications, list the name of all medications, where you are getting them, where study medications will be stored during study, dose range (justify), duration of drug use, labeling of medication, antidotes that must be available, disposal of unused medications, dose schedule (justify), how administered, and potential side effects. If you do not have prescription privileges at the study site one of your co-investigators must be the person controlling use of the drugs and tracking the patients. If you are going to pay subjects to participate, the payment must not exceed the reasonable cost to participation (e.g., transportation) or the payment is considered to be coercive in recruiting subjects.*

(d) Evaluations made during and following project: *This includes medical exams made after your intervention, follow-up plans, etc.*

(e) Risks to subjects and precautions that need to be taken for patient safety: None is not an adequate response. ***NOTE: IT IS YOUR RESPONSIBILITY TO BE CERTAIN THAT APPROPRIATE STUDY INSURANCE IS IN PLACE IN THE EVEN A SUBJECT IS INJURED DURING PARTICIPATION. YOU CAN BE SUED SUCCESSFULLY IF A SUBJECT FILLING OUT A SURVEY GETS A PAPER CUT.***

(f) Method of data analysis: *This is a very detailed section. Every hypothesis and objective has a fully detailed method of data analysis. You must include which tests you propose to use and why. If you did a two group study in which you saw each subject several times, saying that you will use an analysis of variance is not sufficient. You need to say that you will use a repeated measures analysis of variance with *x* being the repeated measures and *y* being the... You must also identify which variables are suitable for parametric or non-parametric statistics and insure that you have chosen the correct tests for each type of variable. Be sure to include both graphic and inferential analyses.*

(g) Roles of the investigators: *Include if not described previously. Include % time on project.*

12. Resources and budget: *Detail exactly what you need to do the project. Where are you getting the equipment and other materials you need to do the project? Who is paying? Are there mailing costs involved? If you are being funded by a source outside of your institution, you must have a letter from your institution's Director of Research stating that the funds are acceptable.*

13. References: *These are presented in detail (names of all authors, title of article, full title of journal, volume and pages, and year of publication) using the format of the journal to which you are most likely to submit your findings. Use the APA's format if in doubt.*

14. Your signature and signature block as well as (if you are a student) that of your thesis chair follows this statement:

As principal investigator, I will ensure that the clinical investigation has been approved by the proper review committee(s) before starting, changing, or extending the investigation. I will ensure that appropriate liability insure covering this protocol is in place prior to starting it. I will also ensure that all subjects or their representatives, including subjects used as controls, are fully informed of the nature of the investigation to include potential risks to the subject.

15. Impact statements: *Each group involved with your study must agree that you can use their*

time, resources, people, space, etc. Place letters from each here.

16. Curriculum Vitae: Place CVs for everyone involved in your study here. This includes all co-investigators and consultants. Use NIH study CVs which are limited to a maximum of three pages. The first page has your educational and work history and, if space permits, publications. You can use up to two additional pages for publications. The idea is to demonstrate that the investigators have sufficient background and expertise to tackle the proposed project.

17. Study Explanation / Agreement to Participate: If you are not conducting the study at the institution approving the study and your home institution requires their own agreement, use theirs, not ours. Put the document after this page. A sample of a typical agreement, and instructions for using it, begins on the next page.

Study Explanation / agreement

Ground Rules:

1. This document assures that participating subjects have had an adequate explanation of the study and helps protect you and the institution if problems arise.
2. They protect the investigator and the host institution in the event of later disagreements.
3. They must be written so an average sixth grader can actually fully understand the study and the requested participation.
4. You must include a brief explanation of why the study is being done, a complete explanation of what the subject will be asked to do, a description of all tests the subject will participate in, an explanation of what will happen to any data gathered (e.g. will the results of a diagnosis providing psychology test be given to officials), privacy considerations, alternatives to participation, consequences of not participating, risks, and inconveniences.
5. Minors mentally able to understand the study must freely assent to participation by signing the consent form. A legal guardian must consent to the minor's participation by signing the consent form also. Minors too young to fully understand the study must be given as full an explanation of the study as they can grasp. If they don't want to participate, you must not be a party to coercing their participation.
6. People need to know that they do not have to participate in your study and that they can drop out whenever they wish without prejudice but that they may have to have an exit exam if they are under treatment.
7. A typical study explanation / agreement to participate follows.

Study explanation / agreement to participate In a Study Authorized by the Sample Institutes

Study title and approval date here in large letters

I, _____ SSN _____
having full capacity to consent and having attained my _____ birthday, do hereby volunteer to participate in the research protocol **type in protocol title** under the direction of **type in name of principal investigator** conducted at **type in place study will be conducted**.

The implications of my voluntary participation; the nature, duration and purpose of the research study; the methods and means by which it is to be conducted; and the inconveniences and hazards that may reasonably be expected have been explained to me by _____.

I have been given an opportunity to ask questions concerning this investigational study. Any such questions were answered to my full and complete satisfaction. **Should any further**

questions arise concerning my rights, I may contact the Director Research at the Sample Institute, (206) ____ - ____.

I understand that I may at any time during the course of this study revoke my consent and withdraw from the study without further penalty or loss of benefits. My refusal to participate will involve no penalty or loss of benefits to which I am otherwise entitled.

PART B - EXPLANATION OF WHAT IS TO BE DONE

(Please start with this paragraph) INTRODUCTION: You have been invited to participate in a clinical research study conducted at _____. Participation is entirely voluntary; refusal to participate will involve no penalty or loss of benefits to which you are otherwise entitled.

PURPOSE: state the purpose of the study in lay terms

PROCEDURES: List the details of the experimental treatment, procedures here; if routine and experimental procedures are both to be done, differentiate between what is routine and what is experimental; if drugs are used state the route and dosage; if experimental drugs are used use the phrase: "This study involves the use of an investigational drug called **insert drug name**. This means that the drug has not yet been approved by the Food and Drug Administration (FDA) for widespread use, but the FDA has agreed to its use in this study of the safety and effectiveness in treating/preventing/diagnosing (**indication**).

POTENTIAL BENEFITS: Benefits expected for subject - be very specific.

RISKS, INCONVENIENCES, AND DISCOMFORTS: Risks, inconveniences, and discomforts that may occur during the subjects participation; don't forget risks of extra blood draws or x-rays. Include extra trips to the research location and longer sessions as inconveniences.

If blood or tissue is obtained add: "The blood/tissue samples (**specify type of sample**) you are providing may also be used in other research studies. You will not be personally identified in any publication of the results of these other research studies. In addition, you agree to waive all property rights to these (**enter type**) samples.

ALTERNATIVES TO PARTICIPATION: If this is a treatment protocol, explain all standard treatments that would be available to the patient if he/she were not on the study.

The following are required paragraphs; modify if necessary.

CONFIDENTIALITY OF RECORDS: The case records from this study will be available for review by members of the Human Use Committee at _____ (your institution) and possibly by representatives of the Food and Drug Administration. Otherwise, only the people conducting this

study will have access to the records from this study. Information gained from this study may be used as part of a scientific publication, but you will in no way be personally identified.

OTHER INFORMATION: Significant findings that occur during this study which might affect your decision to participate in the study will be discussed with you. Any significant findings developed from this study will be available to you and may be obtained from the investigator. Your participation in this study may be terminated without your consent if conditions occur which might make your continued participation dangerous or detrimental to your health. The study itself may be terminated prior to your completing participation.

If you should require medical care for injuries or disease which result from participation in this study, **fill in the legal arrangements you have made. It is your responsibility to make certain that the institute has appropriate insurance in place for your study.**

You are encouraged to ask any questions, at any time, that will help you to understand how this study will be performed and/or how it will affect you. You may contact **name of PI at telephone number** for further information.

IF THERE IS ANY PORTION OF THIS EXPLANATION THAT YOU DO NOT UNDERSTAND, ASK THE INVESTIGATOR BEFORE AGREEING TO PARTICIPATE IN THIS STUDY. You will be given a copy of this consent document for your records.

I do ~ do not ~ (check one & initial) consent to the inclusion of this form in my medical treatment record.

SIGNATURE OF VOLUNTEER	DATE	SIGNATURE OF LEGAL GUARDIAN (if required)
PERMANENT ADDRESS OF VOLUNTEER	TYPED NAME OF WITNESS	
	SIGNATURE OF WITNESS	DATE SIGNED

Section H

Examples of protocols, grants, and papers

Sample 1

Protocol and consent form for a **pilot** study

**Human Use Protocol
Sample Institutes**

1. **Date submitted to the Human Use Committee:** 20 February 1997
2. **Title:** Treatment of aura inaugurated migraine headaches (classic migraine) with pulsing electromagnetic fields: A pilot efficacy study
3. **Investigators:**

a. **Principal Investigator:** Linda A. Example, BA.
Marktown, TR (372) 923 - 8877
Graduate student at Sample Institute

b. **Associate Investigators:**
James E. Expert, PhD
Hometown, WA (206) 819 - 6423
Staff, Sample Institute

Lucey Nervesplit, MD
Neurological Associates, Inc.
Saffron, WA (360) 149 - 3948

4. **Summary:** This pilot study is intended to determine whether Pulsing Electromagnetic Field (PEMF) therapy has any clinical efficacy in treating classic migraine headaches. The study is being performed because a patient being treated with PEMF for a non-union fracture (which is one of the standard uses of the device) resulted in the unexpected side effect of eliminating the patient's headaches. The patient had a history of having classic migraines preceded by a visual prodroma at least twice a week for many years. The headaches, but not the prodroma, stopped when PEMF therapy was initiated and did not return either during the six weeks of treatments (five times per week) or for the two months since the treatment ended upon successful union of the fracture. Our team subsequently phoned all twenty-one female subjects who had completed participation in a pelvic stress fracture - PEMF treatment study. Two of the nineteen subjects who could be located reported that they had histories of migraine headaches. Both reported that their headaches had decreased or stopped during or after participation in the study.

PEMF exposure has been shown to increase peripheral blood flow in the limb it is applied to. Other successful therapeutic approaches to migraine headaches such as temperature biofeedback also cause increases in peripheral blood flow. Thus, if PEMF has any effect at all, it may be through the same mechanism. In over thirty years of use, no significant side effects have been reported from the use of PEMF generators so the exposures are not likely to pose any significant risk.

We propose to have ten adult patients of either sex between the ages of 18 and 70 with at least a two year history of having classic migraines at least once per week keep a daily log of the frequency and intensity of headaches as well as medication use for one month. They will then be exposed to PEMF on the thigh at a power/frequency setting of 6/600 for one hour per day, five days per week for three weeks. This should be more than sufficient time to produce any effect. They will continue keeping the headache log during the PEMF period and for two months after PEMF exposure.

Treatment success is usually defined as at least a 50% decrease in headache activity as calculated from a composite score based on frequency, duration, and intensity with a commensurate decrease in medication use. The intervention will be considered successful if at least half of the subjects reach this criterion.

5. Facilities to be used: Sample Institute's research facility. The director of Research has provided a letter agreeing to this use.

6. Time course of study:

- a. Anticipated start date: 2 April, 1995
- b. Anticipated completion date: 30 September, 1995

7. Hypothesis: That application of PEMF to the inner thighs of migraine headache sufferers will result in decreased headache activity.

8. Objective: To determine whether classic migraine headaches can be treated with pulsing electromagnetic field (PEMF) therapy. This pilot will only determine whether the application of PEMF appears to have a clinically important effect. If it appears to have such an effect, larger, controlled studies will be proposed which will determine the extent and duration of the effect.

9. Medical application / significance to the field: Most adults in the United States have at least occasional headaches. As there is no longer an accepted way to differentiate between migraine and tension type headaches, and almost all have components of both (so are termed mixed headaches), there is no way to determine the impact of one particular type of headache - if there actually are different types. Headache is now the leading medical cause of lost days of work and costs the U.S. many billions per year to treat. The Nuprin Pain Report (1987) found that 157 million work days per year were lost due to this problem alone. A large minority of patients with migraine headaches are not adequately controlled with current treatments. Many of the effective treatments have significant side effects and require life-long drug therapy. This study is a direct effort to determine whether a technique already in use for other problems can effect migraine headaches as well. If it can, a new technique would be added to the currently flawed armamentarium. Its demonstrated lack of side effects makes it especially attractive.

Most theories concerning the underlying mechanisms producing migraine headaches have collapsed under the weight of negative evidence. However, it is known that effective treatments such as temperature biofeedback for warming the fingers do cause an increase in peripheral vasodilation. Recent studies on Raynaud's disease have shown that this effect is not mediated by the sympathetic nervous system but, rather, by altering circulating hormones whose release is under CNS control. Thus, it is not known whether temperature biofeedback works through peripheral or central mechanisms. The same would be said of peripheral vasodilators. PEMF causes an increase in peripheral vasodilation through peripheral mechanisms. If it has a similar effect on migraine headaches as temperature biofeedback, then we will have some hints about physiological changes which alter migraines.

10. Status:

a. Incidence, impact, and request for treatment of migraine and tension headaches:

Most adults in the United States have at least occasional headaches. Headache is now the leading medical cause of lost days of work and costs the U.S. many billions per year to treat. The Nuprin Pain Report (1987) found that 157 million work days per year were lost due to this problem alone. Numerous surveys of the general population have indicated that about 65 percent of males and 78 percent of females reported having had at least one headache within the past year (Linnet et al 1989, Pascual et al 1990). About half of the men and 65 percent of women report having at least one headache per month. About 30 percent of males and 44 percent of females reported that these were severe with about 15 percent having headaches severe enough to affect daily activities. The data hold for young adults of typical military age. Among young adults, severe headaches lasted about six hours for males and eight for females with eight percent of males and fourteen percent of females losing a day or more of work per month due to headaches. About six percent of people attending civilian general medical practices request treatment for headaches as their primary reason for coming. About seven percent of males and seventeen percent of females reporting headaches requested treatment within the last year (Blanchard and Andrasik 1985).

b. Relationships between headaches and peripheral bloodflow: The chain of events initiated by increasing peripheral blood flow which result in the sustained decrease in headaches is not understood (Reich and Gottesman, 1993) but it is known that changes in extracranial blood flow do influence headache activity for about half of the migraine headache patients studied (Gillies and Lance, 1993). Most of the evidence giving direct support for relationships between extracranial blood flow in the head and migraine headaches has been discredited in the last few years (Gillies and Lance 1993) and the correlations between intracranial blood flow changes and migraine activity are really restricted to changes related to the aura. Thus, we really know very little about relationships between blood flow and migraine headaches.

Temperature biofeedback from the hands and feet has been shown to increase peripheral blood flow (Freedman 1991, Iezzi, Adams, and Sheck, 1993). Double-blind and clinical studies have shown that this training results decreased migraine headache activity which is sustained for ten to fifteen years (Blanchard and Andrasik 1985, Iezzi, Adams, and Sheck, 1993). The facts that (a) people's peripheral blood flow can increase during biofeedback and that (b) biofeedback can result in decreased headache activity do not mean that increasing peripheral blood flow has any direct bearing on decreasing headache activity. It is entirely possible that both effects are related to some more central mechanism. For example, simply relaxing may result in slightly decreased sympathetic tone, which would result in increased peripheral blood flow. Relaxing while learning to recognize and deal with stressful events could result in a decrease in stress related headaches. Yates (1984) has shown that nine out of ten people not trained in fingertip temperature control can not tell whether their fingertip temperatures are going up or down when they randomly fluctuate by 0.1 degrees.

Heat emanating from the fingertip is virtually entirely caused by blood flowing through the finger. Shunts between the arteries and veins can vary their size rapidly. This variability is largely controlled by the sympathetic nervous system in response to changes in external

temperature. The capillaries can be constricted to a very limited extent by the nervous system but there are **no** nerves which can control dilation of these capillaries in human fingers. Thus, people can not force their fingertip capillaries open.

Freedman (1991) has found that increases in finger blood flow are controlled nearly entirely by a beta-adrenergic vasodilating mechanism. Both alpha and beta adrenergic receptors are located on the blood vessel walls. Circulating catecholamines released from the adrenal medulla and, possibly, from other nerve endings elsewhere in the body clearly cause vasoconstriction. An as yet unidentified substance apparently causes vasodilation when released from other sites. Several investigators (as reviewed by Freedman 1991) have shown that general relaxation accompanied by decreased heart rate and respiration need not be accompanied by increased blood flow to the finger. There is also no evidence that decreased sympathetic tone accompanying relaxation is accompanied by increased blood flow to the finger. Thus, while subjects may use biofeedback devices to monitor their progress in elevating their finger temperatures, it is not likely that either a general relaxation response nor some specific learning related to the finger alone is causing the increase in temperature.

According to Wauquier et al (1995), brain arterial diameter and blood flow is affected by PaCO₂ (partial pressure of carbon dioxide in the cerebral arteries), not by the autonomic nervous system. Transcranial doppler studies show that temperature biofeedback decreases cerebral blood flow velocity in the middle cerebral artery. Similar results were found by Claghorn et al (1981) using a Xenon inhalation technique. They found that people trained to warm their hands had different effects than those trained to cool them and that both were different from people with no histories of migraine headaches. Levine et al (1987) found that asymmetries in cerebral blood flow patterns occur among migraine headache patients even when they are not having headaches. Thus, it is very likely that biofeedback for control of migraine headaches works by changing blood flow in the brain rather than through any effect on the sympathetic nervous system. The only remaining evidence we know of supporting extra-cranial mechanisms is from a study by Feurstein et al (1983) in which the investigators found that temporal arteries dilated starting about three days prior to a migraine headache and constricted the day before it. Neither pain nor anxiety were directly related to dilation of the temporal artery.

Thus, it is likely that increasing peripheral blood flow by any means will result in a decrease in frequency of migraine headaches.

c. Effect of pulsing electromagnetic fields on peripheral blood flow: This technology has been in use since the 1950s. Units of the type we propose to use produce pulsed high-frequency, high peak power electromagnetic energy at a frequency of 27.12 MHZ in 65 microsecond bursts occurring in sequences ranging between 80 and 600 pulses per second. Wattage ranges from 293 to 975 peak watts for some units and less for others. Both pulses per second and wattage can be set in any of six steps. The field extends about 12 cm from the unit's head. The unit's head is placed just above the area to be exposed and turned on for a set amount of time. A typical generator is illustrated in Figure One. Various units differ slightly in a variety of ways such as the exact shape of the wave, rise and fall times, and power output. There is no actual evidence that any of these differences have any clinical importance. Most of the following studies were performed with Diapulse units (Diapulse INC. of New York) so the results may not apply to other devices.

Erdman (1980) recorded peripheral blood flow from twenty normal subjects using both a temperature probe and volumetric measurements while they were being exposed to PEMF generated fields. He found a high correlation between the amount of energy produced by the device and peripheral blood flow with increases beginning within about eight minutes and plateauing by 35 minutes. Pulse rate and rectal temperatures did not change. This relationship has been confirmed in basic studies of blood flow in rabbit ears (Fenn 1969). Ross (1990) recently reviewed the basic science and animal studies as well as some of the clinical studies showing the effectiveness of PEMF generators in increasing blood flow and wound healing. Cameron (1961) demonstrated increased rates of healing in experimentally induced wounds in dogs. Goldin et al (1981) found similar results among humans in a double blind study using changes in fibroblast concentration, fibrin fibers, and collagen in the wound sites and in swelling. The recent status of the emerging field of electromagnetic medicine has been reviewed in a book by O'Connor et al (1990).

Figure One:

Typical Pulsing Electromagnetic Field Generator



d. Visual - analog pain rating scale: The scale is used with our patients to help them rate the intensity of their pain. It consists of a colored bar with numbers under it. The bar is white on the left end and gradually changes through darker shades of pink to deep red at the right end. The numbers go from zero under the white through ten under the deep red. The words "no pain" are printed to the left of the zero and white area of the bar and the words "maximum pain" are printed to the right of the "10" and deep red area of the bar. The patient is shown the scale and told that zero indicates no pain and that ten indicates the most pain imaginable. The patient

looks at the scale and gives a number representative of the pain intensity at that moment or over the time period requested by the interviewer. The visual-analog scale has been in use for many years and has been determined to be the most reliable and valid way to assess changes in an individual patient's pain intensity across numerous evaluation sessions. Reliability and validity have been reviewed by Huskisson (1983). He notes that correlations between successive measurements of pain have been as high as 0.99 and are usually at the level.

e. Evaluation of headache activity: Treatment success is usually defined as at least a 50% decrease in headache activity based on frequency, duration, and intensity with a commensurate decrease in medication use (Blanchard and Andrasik 1985). Subjects in headache studies usually keep a daily log of the frequency, duration, and intensity of headaches as well as use of headache related medications before, during and after the intervention period(s). Subjects usually rate their pain on a visual analog pain scale as discussed above. The efficacy of logs (sometimes called diaries or daily charts) for tracking headache activity is very high (McKee 1993, Blanchard and Andrasik 1985).

Blanchard and Andrasik (1985) reviewed the types of headache logs commonly in use and their validity and reliability. They found that subjects do not keep daily logs requiring several entries per day honestly. Rather, after a week or so, the subjects fill in events from memory. As numerous studies have shown that these memories for pain and related events are flawed to the point of uselessness, there is no point in asking people to keep detailed logs for several months. In one study, they found that only 72% of highly motivated staff members trying to test the validity of the type of log used in their clinic were able to keep a diary requiring four entries per day for two weeks. They found that the logs correlated well with reports from "significant others" about headache complaints and with global ratings of headache activity. Thus, the logs can be valid and reliable when in a useable format.

The type of log proposed for use in this study requires the minimum possible subject compliance while gathering the most crucial data needed to differentiate migraines from other types of headaches (in relationship to knowledge of the subject's headaches ascertained during the in depth interview and medical evaluation). It only requires subjects to make one entry after each headache. They only have to enter the headache's date of occurrence, duration, presence or absence or an aura, vomiting, location of pain, worst and average intensity, and medications or other interventions utilized. Thus, subjects having to keep a log for months do not have to keep making endless notations in a log. We have found over 18 years of having research subjects keep progressively simpler pain logs, that this is the most reliable format for long term use. We have validated it using telephone questioning and in person interviews. Although we never published the data, nine years ago, we went over several hundred "multi-entry per day" logs kept for a previous headache study and found that the vast majority had obviously been filled out for most of the week at one sitting as the pen used was the same and the repetitious form of the numbers (mostly zeros) running down columns (instead of across) was obviously just a repetition. When we repeated this informal test last year with the current version of the log, it was equally obvious that people were making separate entries. A shortened version of the proposed log is presented in Figure Two. Andrasik (1992) has recently endorsed this approach for the type of discontinuous headaches we are working with and used it successfully in at least

one of their studies (Andrasik et al 1985).

11. Study Plan:

a. Subjects:

(1) Number of subjects: Ten. The most common way to determine whether a headache treatment is efficacious is to determine whether at least half of the patients have a fifty percent decrease in a composite headache activity score calculated from headache intensity, duration, and frequency, and a corresponding decrease in use of headache related medications. We can not begin to guess at this with less than ten patients but do not want to include more because this is just a pilot study to get an idea of whether the technique has any value at all. We anticipate that several subjects will fail to return the follow-up logs so as many as twelve subjects may have to begin the study for us to have ten complete data sets.

(2) Age range of subjects: 18 - 70. People over seventy typically get less migraines than most adults. People under 18 are considered minors and this protocol is designed for adults to limit variability.

(3) Sex and racial composition of subjects: No restrictions.

(4) Source and availability of subjects: Subjects answering advertisements who meet the entrance criteria and can come to Sample Institute five days per week for three consecutive weeks.

Figure 2: **Typical log for discontinuous headache**

Headache Log

Your Name: _____

Date you started this

log: _____

day month year

Date you ended this

log: _____

Please fill in this log every day that you have a headache. Keep the log for at least **FOUR WEEKS**:

Date	Pre-Headache Aura (lights, smells)	Symptoms during headache					# days since start of last menstrual cycle	Duration (hours)	I (0
day/mo/yr	Yes or No. If yes, describe.	Location of headache	Feeling: Pulsing? Dull?	Vomit? yes/ no	Light / sound sensitive?	Neck or shoulders hurt?		Avera	

Rate the intensity of your headache on a scale of zero through ten in which zero is no headache and ten is the most severe you can imagine - you would not be able to bear it for one more second without fainting. You will not make an entry if your headache would be rated at zero (no pain). Many people break out the numbers between one and ten as follows:

1-2 = Mild

3-5 = Moderate

9 = Excruciating

6-8 = Severe

10 = So severe that you would pass out if

you had to bear it for one more second

(5) Inclusion, exclusion, and diagnostic criteria: Patients of either sex eligible for care at Sample University Clinic between the ages of 18 and 70 who have at least a two year history of classic migraine headaches with prodromas at least once per week but no other major medical or psychological problems. Subjects must have been diagnosed as having uncomplicated classic migraine headaches by a neurologist in order to participate. Headache diagnoses will all be confirmed by the participating neurologist according to data gathered during an initial interview according to the standard International Ad Hoc Committee classification (Olesen 1994). In order to participate, patients must have some form of pre-headache prodroma /aura (such as changes in visual, tactile, auditory, taste, etc. sensations), and either (a) vomit during the headache (with temporary pain relief) or (b) have pulsing pain. No patients with traumatic onset, medicine rebound, sinus, cluster, tension or other types of headaches will be able to participate.

Although there is no history of pregnancy related complications arising from use of the device by pregnant women, the manufacturer includes a caution statement with the device and no studies have actually ruled out pregnancy related problems. Thus, we will exclude pregnant women just to be on the ultra-conservative safe side. Women must have a urine test showing that they are not pregnant prior to beginning participation and must agree to use a method of birth control which is at least 85% effective during the study. If a participant becomes pregnant during the study they will have to drop out immediately. History of other kinds of headaches or headaches more than five times per week are exclusion criteria.

(6) Identification of subjects / privacy: Each subject's data will be given a sequential group code when stored outside of the medical record. Clinical records will be kept in the usual way. Additional information recorded for study purposes will be kept in a locked file until patient identification is removed and coding is substituted.

b. Evaluations before entry: Each subject will already have been evaluated by a neurologist prior to being interviewed for the study. Each subject will be interviewed by one of the investigators to insure that the subject's headaches meet the entrance criteria. The format for the structured interview is described in Figure Three.

Figure Three

Structured interview for prospective subjects

1. Confirm contact phone numbers.
2. Review history of headaches (when began, traumatic vs. unknown, vs. puberty, etc. origins)
2. Elicit details of neurologist's diagnosis of migraines. Copy relevant portion of medical record.

3. Review temporal pattern of migraines - duration of each headache, frequency per month, spread throughout year, relationship to menstrual cycle, etc.
4. Review history and current treatments - be very specific about dosage/timing of drugs and check apparent quality of behavioral interventions.
5. Elicit details of pattern of headache onset - timing, diet, sleep, hydration, precipitating events, stress, etc.
6. Elicit details of prodroma - e.g. confusion, visual field, auditory, olfactory, tactile, clumsiness, etc.
7. Elicit history of events during the headache - typical location (s), concurrent shoulder/neck pain, vomiting that alleviates pain, pattern of pain (steady, pulsing, millisecond spikes, etc.), photophobia, etc.
8. If clearly not migraine, traumatic origin, or complex, dismiss subject. Otherwise, forward results to neurologist.
9. Set date to begin exposures.

c. Methods: Subjects will keep the log of headache activity detailed above for one month prior to participation. As discussed above, this should be a sufficient duration to get a reasonable idea of headache patterns (Blanchard et al, 19887). Subjects will rate their pain on the visual analog pain scale also discussed in the introduction. The scale goes from zero (no pain) to ten (so much pain that they would faint if they had to sustain it for one more second).

After completion of the baseline, subjects will be exposed to PEMF on the inner thigh at a power/frequency setting of 6/600 for one hour per day (30 minutes per thigh) , five days per week for three weeks. This is the highest setting the machine provides. It will be used because previous studies (discussed in the introduction) found a positive correlation between the field's power and amount of increase in blood flow. As we wish to maximize increased peripheral blood flow and no problems have been reported in over 35 years of using the device at maximum power, this is the setting to start at. The inner thigh will be exposed because this is the site that worked with the trial patient and because the major blood vessels to the leg run under the site. The necessity for daily exposures is derived from our experience with diabetic ulcer patients and non-union fracture who heal faster if exposures are daily (with as few skips as possible) and for at least an hour per day. This should be more than sufficient time to produce any effect. Subjects will continue keeping the headache log during the PEMF period.

At the end of the two weeks of exposure to PEMFs, subjects will keep their logs for at least one more month. Although not a formal part of this pilot study, if there are any effects on headache activity what-so-ever, we will ask the participants to continue keeping their logs indefinitely so the time course for any effects which may occur can be evaluated.

We anticipate being able to recruit about two patients per week who will meet the study criteria. It is likely that several people will drop out by not sending in their follow-up logs so a total of twelve people will probably begin the study. As each subject will require about three months to participate, about five and a half to six months will be required for all subjects to complete participation.

d. Evaluations during and after participation: The only evaluations will be the headache log kept for one month before intervention, during the three weeks of intervention, and for one month after intervention.

e. Risks: There are no known risks to participation.

f. Statistical considerations: Analysis of headache activity will be performed by making a composite rating for each subject for each of the three rated periods (before, during, and after intervention). Activity for each period will be calculated for each variable by simply adding up the ratings (e.g. total hours of pain for the period) and by constructing a composite score equal to frequency times intensity for each period. The parametric measures (e.g. hours of pain) will be compared using a parametric one way, repeated measures analysis of variance while the non-parametric measures (e.g. pain intensity) will be evaluated using the non-parametric equivalent.

g. Roles of investigators:

(1) Linda A. Example, BA, is the project's principal investigator and major worker. She will recruit all the subjects, conduct the screening interviews, schedule the subjects for PEMF exposure and perform the exposures, gather the log data, reduce the data, and participate in the data analysis. She will spend about $\frac{1}{3}$ of her time on the project.

(2) James E. Expert, PhD, is a psychophysicologist with considerable experience in the use of pulsing electromagnetic fields for treatment of pain as well as in evaluation and analysis of headache data. He will provide ongoing guidance concerning the use of the PEMF generator and will guide the principal investigator's analysis of the data. He will spend about one percent of his time on this project.

(3) Lucey Nervesplit, MD is a neurologist with experience in the diagnosis and treatment of headaches. She will review the interview data and abstracts of the subjects' medical records to insure that each meets the criteria for the study. She will also be available to answer subjects' and investigators' questions about changes in headaches as the study progresses. She will assist in reducing the data by interpreting changes in headache symptoms. She will spend a total of about twenty hours on this project.

12. Resources and budget: This pilot will be performed entirely with internal resources. The device required to perform the study is already available and the labor will be provided by the principal investigator.

13. References:

- Blanchard, E and Andrasik, F: Management of chronic headaches. Pergamon Press, NY, 1985
- Blanchard, E, Hillhouse J, Appelbaum K, Jaccard J: What is an adequate length of baseline in research and clinical practice with chronic headache? Biofeedback and Self-Regulation 12(4) 323-329, 1987.
- Cameron B: Experimental acceleration of wound healing. Am J of Orthoped 3: 336 - 343, 1961.
- Claghorn J, Mathew R, Largen J, Meyer J: Directional effects of skin temperature self-regulation on regional cerebral blood flow in normal subjects and migraine patients. Am. J. Psychiatry 138: 1182 - 1187, 1981.
- Erdman, W: Peripheral blood flow measurements during application of pulsed high frequency currents. Am J of Orthopedics 2: 196-197, 1960.
- Fenn, J: Effect of PEMFs (Diapulse) on experimental hematomas. Canadian Medical Association Journal 100: 251- 254, 1969.
- Freedman, R: Physiological mechanisms of temperature biofeedback. Biofeedback and Self-Regulation 16: 95 - 115, 1991.
- Fuerstein M, Bortolussi L, Houle M, Labbe E: Stress, temporal artery activity, and pain in migraine headache: A prospective analysis. Headache 23: 296 - 304, 1983.
- Gillies J.D. and Lance J.W.: Pathophysiology of migraine. Chapter ten in (C Tollison and R Kunkel, eds.) Headache: Diagnosis and Treatment. Williams and Wilkins, Baltimore, 1993.
- Goldin, J., Broadbent, N., Nancarrow, J., and Marshall, T.: The effects of Diapulse on the healing of wounds: a double-blind, randomized controlled trial in man. Brit J of Plastic Surg 34: 267 - 270, 1981.
- Huskisson, E : Visual Analog Scales. Chapter in Pain Measurement and Assessment edited by R. Melzack, Raven Press, NY 1983.
- Iezzi A, Adams H, and Sheck C: Biofeedback and psychological management of migraine. Chapter 14 in (C Tollison and R Kunkel, eds.) Headache: Diagnosis and Treatment. Williams and Wilkins, Baltimore, 1993.
- Levine S, Welch K, Ewing J, Robertson W: Asymmetric cerebral blood flow patterns in migraine. Cephalalgia 7: 245 - 248, 1987.
- Linnet, M., Stewart, W., Celentano, D., Ziegler, D., and Sprecher, M.: An epidemiologic study of headache among young adults. JAMA 261: 2211 - 2216, 1989.
- McKee, M: Headache Diary. Chapter 39 in (C Tollison and R Kunkel, eds.) Headache:

16. Curriculum Vitae: Attached (*Not actually included - the CVS would have shown that the investigators were sufficiently knowledgeable to attempt the work.*)

17. Study Explanation / Agreement to Participate: Attached.

**Study explanation / agreement to participate
In a Study Authorized by the Sample Institutes**

Treatment of aura inaugurated migraine headaches (classic migraine) with pulsing electromagnetic fields: A pilot efficacy study

I, _____ SSN _____
having full capacity to consent and having attained my _____ birthday, do hereby volunteer to participate in the research protocol entitled Treatment of aura inaugurated migraine headaches (classic migraine) with pulsing electromagnetic fields: A pilot efficacy study under the direction of Linda A. Example, BA. conducted at the Sample Institutes of Hometown Island Washington. The implications of my voluntary participation; the nature, duration and purpose of the research study; the methods and means by which it is to be conducted; and the inconveniences and hazards that may reasonably be expected have been explained to me by _____.

I have been given an opportunity to ask questions concerning this investigational study. Any such questions were answered to my full and complete satisfaction. Should any further questions arise concerning my rights, I may contact the Director Research at the Sample Institute, (206) 780-5500. For medical questions concerning my participation I can contact Lucey Nervesplit, MD at (360) 149 - 3948.

I understand that I may at any time during the course of this study revoke my consent and withdraw from the study without further penalty or loss of benefits. My refusal to participate will involve no penalty or loss of benefits to which I am otherwise entitled.

PART B - EXPLANATION OF WHAT IS TO BE DONE

INTRODUCTION: You have been invited to participate in a clinical research study conducted at the Sample Institutes. Participation is entirely voluntary; refusal to participate will involve no penalty or loss of benefits to which you are otherwise entitled.

PURPOSE: You are one of 10 people with migraine headaches who are being asked to participate in this research study which is trying to find out if exposure to pulsating electromagnetic fields (PEMFs) has any effect on migraine headaches.

Pulsing electromagnetic field units have been in use for over thirty years to treat swelling and fractures. They are safe, standard devices and the US Food and Drug Administration (FDA) permits their marketing for these uses. There have been no reports of significant side effects during the entire time they have been in use. The fields are essentially radio waves which turn on and off so fast that they do not heat up whatever they are pointed at. They are generated by a device that looks and sounds like an old fashioned hair-dryer. The cone shaped field extends about one foot from the device's "head" so only the part of you the head is pointed at is exposed to the field.

PEMFs have been shown to increase blood flow to the areas they are pointed at. Several moderately successful treatments for migraine headache appear to work, at least partially, by increasing blood flow to the arms and legs. Knowing this fact, it is possible that pointing the field at your thigh will increase blood flow throughout the body sufficiently to prevent migraine headaches. We saw this happen in one patient who was being treated with PEMFs for a broken bone in her thigh. We want to try it with ten more patients to see whether it actually works.

PRECAUTION: Nobody knows whether pulsing electromagnetic fields pose any risks during pregnancy. In over thirty years of use, no problems with pregnant women have been reported but no formal studies of the effects of these fields have been conducted. Thus, women who can get pregnant will be required to take a urine pregnancy test and must promise to use an effective method of birth control during this study. You must inform your therapist and drop out immediately if you become pregnant.

PROCEDURES: You will be able to continue, but not change, your current treatments if you decide to participate in this study. You will keep a log of your headache activity for one month before being exposed to PEMFs. Your PEMF treatment will consist of being exposed to PEMFs for one hour per day for five days per week for three weeks while sitting in a chair with the "head" of the generator pointed at your upper thigh (half an hour on each thigh). You cannot feel anything when the generator is running. You will keep your log of headache activity throughout the three weeks of exposure to PEMFs and for at least one month afterwards.

POTENTIAL BENEFITS: Your headaches may be blocked or decreased by exposure to PEMF's. If they are, we do not know whether the effects will continue after the end of treatment.

RISKS, INCONVENIENCES, AND DISCOMFORTS: You must come to the clinic five times per week for three weeks. The only other inconvenience is that you will not be able to change your medications as long as you are in the study. If you change your medications, you need to leave the study. Pulsing electromagnetic fields of the type we will use are produced by a safe, standard medical device which has been shown to effectively reduce ankle swelling after strains. You cannot feel the device while it is working and we have no reason to think that there is any danger to you from the PEMFs. There are no special precautions or known risks related to this study. No side effects have been reported in over thirty years of using the device. If we discover any problems with using the device or ways to make it more effective during your participation, we will inform you as soon as we can.

ALTERNATIVES TO PARTICIPATION: If you do not participate in this study, you will receive the standard treatments you would have received in any case.

CONFIDENTIALITY OF RECORDS: The case records from this study will be available for review by members of the Human Use Committee at Sample Institute and possibly by representatives of the Food and Drug Administration. Otherwise, only the people conducting this study will have access to the records from this study. Information gained from this study may be used as part of a scientific publication, but you will in no way be personally identified.

OTHER INFORMATION: Significant findings that occur during this study which might affect your decision to participate in the study will be discussed with you. Any significant findings

developed from this study will be available to you and may be obtained from the investigator. Your participation in this study may be terminated without your consent if conditions occur which might make your continued participation dangerous or detrimental to your health. The study itself may be terminated prior to your completing participation.

If you should require medical care for injuries or disease which result from participation in this study, your treatment will be arranged by the Director of Research for Sample Institute.

You are encouraged to ask any questions, at any time, that will help you to understand how this study will be performed and/or how it will affect you. You may contact Linda A. Example, BA. at (372) 923 - 8877 for further information.

IF THERE IS ANY PORTION OF THIS EXPLANATION THAT YOU DO NOT UNDERSTAND, ASK THE INVESTIGATOR BEFORE AGREEING TO PARTICIPATE IN THIS STUDY. You will be given a copy of this consent document for your records.

I do ~ do not ~ (check one & initial) consent to the inclusion of this form in my medical treatment record.

SIGNATURE OF VOLUNTEER	DATE	SIGNATURE OF LEGAL GUARDIAN (if required)
PERMANENT ADDRESS OF VOLUNTEER	TYPED NAME OF WITNESS SIGNATURE OF WITNESS DATE SIGNED	

Sample 2

Protocol and consent form for a **controlled** study

*NOTE: THE CONTROLLED STUDY IS AN EXTENSION
OF THE ABOVE PILOT STUDY SO ONLY THE
CHANGED PORTIONS ARE INCLUDED*

Human Use Protocol
Sample Institutes

1. **Date submitted to the Human Use Committee:** 10 January, 1998

2. **Title:** Treatment of aura inaugurated migraine headaches (classic migraine) with pulsing electromagnetic fields: A double blind, placebo controlled study

3. **Investigators:**
 - a. **Principal Investigator:** Linda A. Example, BA.
Marktown, TR (372) 923 - 8877
Graduate student at Sample Institute.

 - b. **Associate Investigators:**
James E. Expert, PhD
Hometown, WA (206) 819 - 6423
Staff, Sample Institute

Lucey Nervesplit, MD
Neurological Associates, Inc.
Saffron, WA (360) 149 - 3948

4. **Summary:** An open pilot study showed that headache activity among thirteen patients with migraines decreased from a one month average of 4.03 headaches per week to an average of 0.43

per week after two to three weeks of daily exposure of their thighs to pulsing electromagnetic fields. Follow-ups ranged from one to fourteen months with a mean of eight months. Headaches averaged 0.14 per week during this period. The pulsing electromagnetic field generator was set to produce 975 watt, 27.12 MHZ fields having 65 microsecond bursts pulsing 600 times per second.

The proposed double blind, placebo controlled study will determine whether the results of the pilot were due to placebo effects and whether the effects will last for at least six months. Participants will have at least a two year history of migraine headaches with aura at least three times per month and be of either sex between the ages of 18 and 70. Subjects will keep a daily log of the frequency and intensity of headaches as well as medication use throughout participation. A different therapist will track the subjects and enter their headache data than will perform the exposures in order to keep the study double blind. After a one month baseline, subjects will be randomized into actual or placebo pulsing electromagnetic field exposure groups. They will then be exposed to pulsing electromagnetic fields (real or placebo) on alternating thighs for one hour per day, five days per week for two weeks. At the end of the exposure period, patients will keep the log for six months and will be telephonically encouraged by bi-weekly phone calls. A power analysis of the pilot data indicates that twenty subjects will be required per group at a power of .80 and a significance of 0.05 assuming a thirty percent placebo response and a sustained seventy percent response to actual exposure. Forty-five subjects will begin the study to make-up for anticipated drop-outs.

5. Facilities to be used: Research facility at Sample Institute.

6. Time course of study:

a. Anticipated start date: 30 March, 1998

b. Anticipated completion date: 29 March, 2000

7. Hypotheses:

a. That exposure of patients with chronic classic migraines to PEMF for two weeks will result in at least a 50% decrease in frequency of migraine headaches for one month for at least 80% of the subjects exposed to actual PEMFs and that this difference will be statistically greater than the reduction shown by those exposed to placebo PEMF.

b. That the above reduction can be followed for up to six months to determine the slope of return to baseline levels of headache activity.

8. Objectives: To determine whether exposure to PEMFs results in a decrease in headache frequency relative to exposure to placebo PEMFs and to determine how long (up to six months) the effect lasts.

9. Medical application / significance to the field: *As in the initial protocol.*

10. Status: *As in the initial protocol but add one section on the need for placebo groups and one on the results of the pilot study as follows.*

f. Requirement for a placebo control group: Pain is very reactive to placebo intervention (e.g. Beecher's pioneering study in 1955) so realistic placebos are a requisite part of evaluating any new intervention (Cooper 1981). A placebo / non-specific effects control group is vital to the study design because headache studies usually find about a thirty percent, short term response to inactive interventions. For example, Couch (1993) reviewed twelve placebo controlled headache studies and found a range of placebo response from four to fifty-five percent with most in the thirty percent range. While most studies, including those reviewed by Couch, use medicinal placebos, machines have been shown to produce effective placebo responses as well (Schwitzgebel and Traugott, 1968). Our study would be especially likely to produce a placebo response because of the impressive nature of the device itself and the intense □treatment□ regime which requires patients to make twenty visits to a major medical center. Non-specific effects would also be highly likely as all participants take time out of their normal routines to sit quietly in a comfortable room away from their daily stresses for an hour per day.

g. Results of the pilot study: The pulsing electromagnetic field generator described above (Diapulse model 103) was set to produce 975 watts at 27.12 MHZ in 65 microsecond bursts pulsing 600 times per second. The head of the device (illustrated in Figure 1) was set so that its cone shaped field was pointed at the inner thigh. This was done because a patient being treated for non-unions reported that her migraines had stopped during treatment after her thigh was exposed to pulsing electromagnetic fields for one hour per day for week.

The open pilot study was performed with 13 migraine headache patients to determine whether the decrease in headaches reported by the above patient was just a coincidence. The original pilot was to have only ten subjects but less subjects dropped out than anticipated so a total of thirteen subjects completed the study out of 15 who started it. The two drop-outs did not complete the exposure period. The participants kept a two week, pre-treatment daily headache log followed by two to three weeks of exposure to pulsing electromagnetic fields for one hour per day for five days per week to the thigh. This was followed by a two week post-treatment daily headache log and follow-up phone calls for up to 14 months.

Pre-treatment headache activity averaged 4.03 (+/- 2.02) headaches/wk while post-treatment activity averaged 0.43 (+/- 0.36) headaches/wk. During the follow-up (1 - 14 months, mean of 8.1 months +/- 3.1), activity averaged 0.14 (+/- 0.08) headaches per week ($p=0.0001$; paired □t□ = 5.8 with 10 DF).

This work was presented at the World Congress on Instant Cures held in Marlboro, Newschotlund in 1997 and has been submitted for publication in the Interplanetary Journal of Scientifically Substantiated Instant Cures (1997).

h. Requirement for further research: Patients with long histories of migraine headaches who were exposed to pulsing electromagnetic fields showed an almost complete

cessation of headache activity during the weeks of exposure and for many months thereafter. In spite of the impressive nature of the PEMF generator, the investigators do not feel that the major effects were due to placebo responses because (a) the participants each had multi-year histories of unsuccessful treatments with numerous highly touted therapeutic approaches, (b) the change in headache activity was much greater than would be anticipated due to a placebo response, (c) headache activity remained reduced for an average of eight months - which is longer than would be expected of a placebo, and (d) the rebound effect from stopping the prophylactic medicines should have hit sometime near the end of "treatment" so would have overwhelmed any placebo effect.

Because of the powerful effects demonstrated in this trial, the investigators feel that it is worth performing a double-blind, controlled study to determine whether this intervention is actually effective.

11. Study Plan:

a. Subjects:

(1) Number of subjects: A power analysis (Glantz, 1993) of the pilot data indicates that 20 subjects will be required per group to give a 0.05 probability with a power of 0.80 assuming a 30 percent placebo response and a 70 percent response to the actual exposure sustained at six months. This is somewhat conservative so should provide more than ample subjects. A total of 45 patients will begin the study to allow for our usual drop-out rate.

(2) Age range of subjects: *As in the initial protocol.*

(3) Sex and racial composition of subjects: *As in the initial protocol.*

(4) Source and availability of subjects: *As in the initial protocol.*

(5) Inclusion, exclusion, and diagnostic criteria: *As in the initial protocol.*

(6) Identification of subjects: *As in the initial protocol.*

b. Evaluations before entry: *As in the initial protocol.*

c. Methods:

(1) Design: The study design is illustrated in Figure Three. This is a typical two group, double blind, placebo controlled study with initial baseline and follow-up periods. Patients will be diagnosed as having uncomplicated migraine headaches with aura and will be randomized into real or placebo exposure groups after keeping a one month initial baseline of headache activity as defined in the introduction. This will be followed by exposure to the pulsing electromagnetic field generator with half the subjects receiving actual exposure to pulsing electromagnetic fields and half receiving placebo exposure for two weeks. Subjects will keep the headache log described in the introduction throughout this period. This is followed by a six month follow-up during which subjects continue keep the log every time they get a headache

with biweekly phone calls reinforcing their continuing to keep the log.

Figure 3

Study Structure

1. Patients are evaluated to (1) eliminate patients with headache causing medical problems and medications and (2) to eliminate all patients who have headaches of traumatic origin and of other types than migraine with and without aura.
 2. Patients keep a four week log of headache intensity, description, duration, and frequency as well as of medication use.
 3. Patients are randomized to actual or placebo pulsing electromagnetic field exposure by a computer generated algorithm which will insure that by the end of the study an even distribution of subjects have been entered into each group (placebo or actual exposure).
 4. Two weeks of daily pulsing electromagnetic field therapy (real or placebo)
 - One set of patients uses device □A□ and one uses device □B□. Only the technician calibrating the device knows which is the placebo - and that technician does not know which group individual patients are in.
 - Power set at 975 peak watts and 600 CPS.
 - Five, one hour sessions per week with pulsing electromagnetic field aimed at the inside of the thigh (over the medial quadriceps targeting the femoral artery).
 - Patients will be strongly encouraged not discuss their headaches with the technician performing the exposures.
 - Patients continue to keep their logs throughout this period.
 5. Patients keep the log for the next six months after exposure. This permits accurate evaluation of headache activity after the two week exposure period. Patients are telephoned every two weeks to encourage them to continue keeping the log.
-

(2) Exposure to pulsing electromagnetic fields or placebo: After keeping the three initial baseline week log, patients will be exposed to pulsing electromagnetic field (real or placebo) on the thigh at a power/frequency setting of 6/600 for one hour per day, five days per week for two weeks. The pilot showed that two weeks are sufficient time to produce any effect likely to occur. We direct pulsing electromagnetic fields to the thigh because (a) this is the site that worked during the pilot study and (b) we have found that we get more increase in peripheral blood flow when that site is used than from any others we have tried. Neither the therapist exposing the subject nor the subject will know which group they are in. This will be accomplished by dedicating two pulsing electromagnetic field machines to the study. One will have the field generator disconnected from the circuit and the field indicator lights on the heads of both machines will be covered with opaque caps. The machines will look and sound the same when functioning. One will be marked "A" and one "B". Only the therapist assigned to calibrate the devices will know which is which. Daily calibration is necessary to be sure the active machine is putting out the correct field strength. Patients can not feel the machine working so they will not be able to tell which group they are in. However, as a check, each will be asked whether they thought they were in the real or placebo group after each stage of the study. This will be done by having each rate how certain they are they received the real treatment on a scale of 0 - 10 where zero is not at all certain and ten is sure they received the real treatment.

We have conducted three placebo controlled studies using this device (stress fractures, sprained ankles, and post-surgical wound healing). All of the participants rated their belief that they had been exposed to the actual device. The odds were just chance of knowing whether they were exposed to the actual device. With two exceptions, none of over 400 subjects and patients have been able to feel the device in operation. There is no sensation during exposure. The two exceptions were both people with upper limb RSD with severely decremented blood flow to the limb. Both felt a minor tingling during exposure when blood flow increased (as documented by concurrent videothermographic evaluations).

(3) Duration of the follow-up: Blanchard et al (1987) did an analysis of the stability of headache activity over time for different headache disorders in order to determine the appropriate duration of baselines for each disorder. They have shown that the duration of logs we propose are sufficient to ascertain the basal level of headache activity among the diagnostic categories of patients we propose to work with. Our pilot data indicate that the level of headache activity will return to pre-treatment levels within six months after the end of actual exposure to pulsing electromagnetic fields. Thus, the results of the follow-up will be used to determine how often brush-up therapy would be required.

(4) Time table for the study: The plan for this two year study requires an eighteen month funded data collection period followed by a six month minimally funded follow-up completion period. The research associate who will assist the principal investigator in performing the study is already on board and trained. The first four funded weeks will be spent recruiting patients, setting up the data base and allocating the study rooms and devices. Each subject will require 30 weeks to participate in the entire study (four week initial baseline, two week exposure, and 24 week follow-up period). We can realistically expect to expose six subjects per day mainly because that is our about two-thirds of our maximum sustainable rate of recruiting subjects from our population (see the prevalence study in the introduction) meeting the entrance criteria who can come to our clinic on a daily basis for two week periods. No subjects

will be exposed for the month following the one month start-up period as the first aliquot of six subjects will be keeping their initial month long log. It will require a minimum of 24 weeks for all 12 aliquots of six subjects to complete their two week exposures. Given the realities of recruiting and running subjects in our environment, this will probably take twice as long as optimal. Thus, we are reasonably confident that the last aliquot of subjects will finish their final log period / follow-up by the end of the first 18 months. The remaining funded period will be used to complete late subjects, complete data entry and reduction, and collect the remaining long term follow-up data. The six month minimally funded tail of the study will be used to complete follow-ups on patients who started late, analyze the data, and write the report.

(d) Evaluations made during and following project: *As in the initial protocol.*

(e) Risks to subjects and precautions that need to be taken for patient safety: *As in the initial protocol.*

(f) Method of data analysis:

(1) Changes in headache activity during the exposure portion of the study: Differences in headache activity (defined above) will be determined from the log kept for one month before intervention, during the two week intervention, and for the first month after intervention. The first analysis will consider this to be a repeated measures design with two groups (the placebo vs. the real exposure). The repeated measure will be the three periods during which logs were kept. A two way repeated measures analysis of variance will be used with a probability of 0.05 being considered significant. The number of headaches per week and headache duration are parametric measures so a parametric test can be used as long as the distributions are normal and variances are similar. Headache intensity is a non-parametric measure so a non-parametric test will be used. A Bonferonni correction will be used when evaluating the significance of differences between individual time periods (e.g. between the initial logs of each group). We will use a one tailed evaluation because we are predicting that exposure to pulsing electromagnetic fields will produce a decrease in headache activity relative to exposure to a placebo device. Thus we are predicting not only a difference but the direction the difference will be in.

We are also interested in establishing the risk of subjects having headaches after exposure to real vs. placebo fields. This will give us the same information as the number of subjects in each headache type group (subdivided by placebo vs. real) who reach the success criterion defined by Blanchard and Andrasik (1985) of a 50% decrease in headache activity. The risk difference for each group will be calculated in accordance with Overvad's (1994) formula.

(2). Evaluation of headache activity for the six months between the last exposure and end of the study: Differences in the number of headaches per month will be analyzed using a two way repeated measures analysis of variance with the repeated measure being each month's activity. The rate of headaches during the initial baseline will be transformed into a monthly rate so return to baseline rate can be calculated as part of the same analysis. A correlation between number of headaches per month and time will be made for each type of headache and differences in slopes of the lines will be calculated to determine differences in rate of return to baseline (if this happens).

(g) Roles of the investigators: *As in the initial protocol.*

12. Resources and budget: This work will be performed using the resources and devices available within the Sample Institute's research facility. A letter from the Director of Research committing to the use of the required resources for one year is attached. All of the work except performing the actual PEMF exposures will be performed by the principal investigator. The technician who will expose the subjects to PEMF will be paid from a grant given by the National Headache Foundation for that purpose.

13. References: *As in the initial protocol but add the references for placebo studies and the source for the power analysis.*

Beecher H: The powerful placebo. JAMA 159: 1602 - 1606, 1955.

Cooper, S: Comparative analgesic efficacies of aspirin and acetaminophen. Archives of Internal Medicine 141: 282 - 285, 1981.

Couch J: Medical management of recurrent headache. Chapter eighteen in (C Tollison and R Kunkel, eds.) Headache: Diagnosis and Treatment. Williams and Wilkins, Baltimore, 1993.

Gillies, J.D. and Lance J.W.: Pathophysiology of migraine. Chapter ten in (C Tollison and R Kunkel, eds.) Headache: Diagnosis and Treatment. Williams and Wilkins, Baltimore, 1993.

Glantz, S: Primer of biostatistics: The program. New York, McGraw-Hill, 1993.

Schwitzgebel R and Traugott M: Initial note on the placebo effect of machines. Behavioral Science 13: 267 - 273, 1968.

14. Investigators' signatures and signature blocks: *As in the initial protocol.*

15. Impact statements: *As in the initial protocol.*

16. Curriculum Vitae: *As in the initial protocol.*

17. Study Explanation / Agreement to Participate: Attached.

**Study explanation / agreement to participate
In a Study Authorized by the Sample Institutes**

Treatment of aura inaugurated migraine headaches (classic migraine) with pulsing electromagnetic fields: A double blind, placebo controlled study

I, _____ SSN _____
having full capacity to consent and having attained my _____ birthday, do hereby volunteer to participate in the research protocol entitled Treatment of aura inaugurated migraine headaches (classic migraine) with pulsing electromagnetic fields: A pilot efficacy study under the direction of Linda A. Example, BA. conducted at the Sample Institutes of Hometown Island Washington. The implications of my voluntary participation; the nature, duration and purpose of the research study; the methods and means by which it is to be conducted; and the inconveniences and hazards that may reasonably be expected have been explained to me by _____.

I have been given an opportunity to ask questions concerning this investigational study. Any such questions were answered to my full and complete satisfaction. Should any further questions arise concerning my rights, I may contact the Director Research at the Sample Institute, (206) 780-5500. For medical questions concerning my participation I can contact Lucey Nervesplit, MD at (360) 149 - 3948.

I understand that I may at any time during the course of this study revoke my consent and withdraw from the study without further penalty or loss of benefits. My refusal to participate will involve no penalty or loss of benefits to which I am otherwise entitled.

PART B - EXPLANATION OF WHAT IS TO BE DONE

INTRODUCTION: You have been invited to participate in a clinical research study conducted at the Sample Institutes. Participation is entirely voluntary; refusal to participate will involve no penalty or loss of benefits to which you are otherwise entitled.

PURPOSE: You are one of 45 people with migraine headaches who are being asked to participate in this research study which is trying to find out if exposure to pulsating electromagnetic fields (PEMFs) has any effect on migraine headaches.

Pulsing electromagnetic field units have been in use for over thirty years to treat swelling and fractures. They are safe, standard devices and the US Food and Drug Administration (FDA) permits their marketing for these uses. There have been no reports of significant side effects during the entire time they have been in use. The fields are essentially radio waves which turn on and off so fast that they do not heat up whatever they are pointed at. They are generated by a device that looks and sounds like an old fashioned hair-dryer. The cone shaped field extends about one foot from the device's "head" so only the part of you the head is pointed at is exposed to the field.

PEMFs have been shown to increase blood flow to the areas they are pointed at. Several moderately successful treatments for migraine headache appear to work, at least partially, by increasing blood flow to the arms and legs. Knowing this fact, it is possible that pointing the field at your thigh will increase blood flow throughout the body sufficiently to prevent migraine headaches. We saw this happen with thirteen patients who participated in an open pilot study. They were exposed to three weeks of the same PEMFs we are using in this study. The frequency of their headaches decreased from 4.3 to 0.16 and stayed down for an average of eight months. We want to try the therapy with 45 more patients to see whether it actually works.

When people with chronic migraine headaches are given treatments known not to work (a placebo) up to one third show at least some temporary decrease in headache activity. Thus, we need to expose some people to a placebo PEMF generator in order to make sure that people exposed to the real generator show greater decreases in headache activity which last longer than those found after exposure to the placebo.

PRECAUTION: Nobody knows whether pulsing electromagnetic fields pose any risks during pregnancy. In over thirty years of use, no problems with pregnant women have been reported but no formal studies of the effects of these fields have been conducted. Thus, women who can get pregnant will be required to take a urine pregnancy test and must promise to use an effective method of birth control during this study. You must inform your therapist and drop out immediately if you become pregnant.

PROCEDURES: You will be able to continue, but not change, your current treatments if you decide to participate in this study. You will keep a log of your headaches for one month before you can begin being exposed to PEMFs. At the end of the one month log you will be randomized (chosen by chance as by the flip of a coin) to receive either real or placebo PEMF. You can not tell the difference between the real and placebo devices as they look, sound, and feel alike. Your PEMF treatment will consist of being exposed to PEMFs for one hour per day for five days per week for two weeks while sitting in a chair with the "head" of the generator pointed at your upper thigh (half an hour on each thigh). You cannot feel anything when the generator is running. At the end of the two weeks of exposure you will continue to keep your headache log for a minimum of six months.

POTENTIAL BENEFITS: Your headaches may be blocked or decreased by exposure to PEMF's. If they are, we do not know whether the effects will continue after the end of treatment.

RISKS, INCONVENIENCES, AND DISCOMFORTS: You must come to the clinic five times per week for two weeks. The only other inconvenience is that you will not be able to change your medications as long as you are in the study. If you change your medications, you need to leave the study. Pulsing electromagnetic fields of the type we will use are produced by a safe, standard medical device which has been shown to effectively reduce ankle swelling after strains. You cannot feel the device while it is working and we have no reason to think that there is any danger to you from the PEMFs. There are no special precautions or known risks related to this study. No side effects have been reported in over thirty years of using the device. If we discover any problems with using the device or ways to make it more effective during you

participation, we will inform you as soon as we can.

ALTERNATIVES TO PARTICIPATION: If you do not participate in this study, you will receive the standard treatments you would have received in any case.

CONFIDENTIALITY OF RECORDS: The case records from this study will be available for review by members of the Human Use Committee at Sample Institute and possibly by representatives of the Food and Drug Administration. Otherwise, only the people conducting this study will have access to the records from this study. Information gained from this study may be used as part of a scientific publication, but you will in no way be personally identified.

OTHER INFORMATION: Significant findings that occur during this study which might affect your decision to participate in the study will be discussed with you. Any significant findings developed from this study will be available to you and may be obtained from the investigator. Your participation in this study may be terminated without your consent if conditions occur which might make your continued participation dangerous or detrimental to your health. The study itself may be terminated prior to your completing participation.

If you should require medical care for injuries or disease which result from participation in this study, your treatment will be arranged by the Director of Research for Sample Institute.

You are encouraged to ask any questions, at any time, that will help you to understand how this study will be performed and/or how it will affect you. You may contact Linda A. Example, BA. at (372) 923 - 8877 for further information.

IF THERE IS ANY PORTION OF THIS EXPLANATION THAT YOU DO NOT UNDERSTAND, ASK THE INVESTIGATOR BEFORE AGREEING TO PARTICIPATE IN THIS STUDY. You will be given a copy of this consent document for your records.

I do ~ do not ~ (check one & initial) consent to the inclusion of this form in my medical treatment record.

SIGNATURE OF VOLUNTEER	DATE	SIGNATURE OF LEGAL GUARDIAN (if required)
PERMANENT ADDRESS OF VOLUNTEER		

Sample 3

Unacceptable protocol and consent form

This protocol is based on the nearly real paper in sample 4. It is an entirely fictitious rendition of what the protocol would have looked like to come up with the resulting paper.

Human Use Protocol
to be conducted at Albright Surgical Center

1. Date submitted to the Human Use Committee: 15 February, 1994

2. Title: Biofeedback therapy for the management of mulehaulers syndrome

3. Principal Investigators: James Martin, M.D.
Larry Owens, Ph.D.
Judy Knowhow, MA, BCIAC
Albright Surgical Center, Cutnose, Georgia

4. Summary: Mulehaulers syndrome is a neurovascular disorder affecting the upper extremities of the body. It is virtually always bilateral with patients reporting chronic, debilitating pain in the shoulders and arms. No treatments are known to be generally effective for this disorder and most patients undergo numerous unsuccessful medical, rehabilitative, and surgical procedures. Biofeedback is known to be effective for many painful disorders but has not been used for mulehaulers in the past.

This study will investigate the effectiveness of biofeedback for the treatment of patients with mulehaulers syndrome. All patients will undergo a thorough diagnostic examination and will be given individualized biofeedback treatment programs.

5. Facilities to be used: Albright Surgical Center

6. Time course of study: About three years

7. Hypothesis: That biofeedback can help patients with mulehaulers syndrome.

8. Objective: To show that patients treated with biofeedback get better.

9. Medical application / significance to the field: There is currently no reasonable way to treat mulehaulers syndrome. If biofeedback helps, a significant addition to the treatment armory will be available and we will be able to bill for our therapy.

10. Status: Mulehaulers syndrome is a neurovascular disorder affecting both upper extremities. It was originally reported among teamsters who had to pull mules forward by their head bridles and is now generally spread through the population of people who pull carts or hold heavy objects in their arms (1). It is virtually always bilateral with patients reporting chronic, debilitating pain in the shoulders and arms (1). The pain is usually stabbing in nature and both range of motion and motor function are frequently impaired. No treatments are known to be generally effective for this disorder and most patients undergo numerous unsuccessful medical, rehabilitative, and surgical procedures (1,2). The disorder appears to be a distant cousin of thoracic outlet syndrome, whose actual etiology and underlying physiological problems are also unclear (3,4,5). It is likely that chronic mechanical pressure on blood vessels and nerves eventually causes the problem in those susceptible to it (4,6). In our hands, nerve conduction studies are frequently positive.

After diagnosis, our normal treatment consists of physical therapy, wrists and elbow splints, lifestyle modification, analgesics, nonsteroidal-anti-inflammatories, muscle relaxants, antidepressants, and nerve blocks. If these are not effective within ten weeks, surgical decompression of the nerves and blood vessels is attempted for selected patients. If all of this fails, the patients are referred for our biofeedback pain management program.

Biofeedback is known to be effective for many painful disorders (3,4,5) but has not been used for mulehaulers in the past. This study was intended to evaluate the contribution of biofeedback to amelioration of the chronic pain and disability among patients diagnosed with mulehaulers syndrome who failed conservative and surgical intervention.

11. Study Plan:

a. Subjects: All patients in our clinic diagnosed as having mulehaulers syndrome will be asked to participate. We will continue entering patients until we get enough to be sure our results are significant.

b. Evaluations before entry: Diagnostic work-up for mulehaulers syndrome.

c. Methods: All of the participating patients will complete our clinic's biofeedback pain management program (a minimum of 15 sessions). The program includes individual training in sEMG and hand temperature biofeedback, cognitive and behavioral techniques, patient education, positive reinforcement, progressive muscle relaxation training, body scanning, autogenic training, visual imagery training, diaphragmatic breathing. The specific training given to each patient is tailored to his or her needs.

Biofeedback will be given using an A&L 663 (Avicron Labs, NJ) with one channel for temperature training and either one or two for sEMG training. Visual feedback in the form of graphs will be provided on color monitors and therapists will give verbal feedback at the end of each session.

After completion of the program, each patient will fill out a post-treatment questionnaire which consists of ten Lacerate scaled questions on (1) the pleasantness of the sessions, (2) the

helpfulness of the therapists, (3) the helpfulness of the biofeedback, (4) decrease in discomfort, (5) the ability to deal with discomfort, (6) ability / stamina to engage in more activities, (7) enjoyment of participation in physical activities, (8) feelings of unhappiness, (9) feelings of nervousness, and (10) amount of perceived control over life.

12. Resources and budget: Only materials already in the clinic will be used.

13. References: (*LEFT OUT ON PURPOSE TO AVOID EMBARRASSMENT*)

14. Signature and signature block:

Judy Knowhow, MA, BCIAC

15. Impact statements: None Required

16. Curriculum Vitae: (*ACTUAL VITAE NOT INCLUDED*)

17. Study Explanation / Agreement to Participate: Attached

Study explanation / agreement to participate

I, _____ SSN _____
having full capacity to consent and having attained my _____ birthday, do hereby volunteer to participate in the research protocol Biofeedback therapy for the management of mulehaulers syndrome under the direction of Dr. Owens conducted at our clinic. The implications of my voluntary participation; the nature, duration and purpose of the research study; the methods and means by which it is to be conducted; and the inconveniences and hazards that may reasonably be expected have been explained to me by _____.

I have been given an opportunity to ask questions concerning this investigational study. Any such questions were answered to my full and complete satisfaction. Should any further questions arise concerning my rights, I may contact the principal investigator.

I understand that I may at any time during the course of this study revoke my consent and withdraw from the study without further penalty or loss of benefits. My refusal to participate will involve no penalty or loss of benefits to which I am otherwise entitled.

PART B - EXPLANATION OF WHAT IS TO BE DONE

INTRODUCTION: You have been invited to participate in a clinical research study conducted at our clinic. Participation is entirely voluntary; refusal to participate will involve no penalty or loss of benefits to which you are otherwise entitled.

PURPOSE: We want to find out if biofeedback cures mulehaulers syndrome.

PROCEDURES: You have been diagnosed as having mulehaulers syndrome. We gave you biofeedback therapy and now we want to find out if it worked. All we want you to do fill in the attached ten question survey about how well your treatment worked.

POTENTIAL BENEFITS: Filling in this survey will help us know how well you liked working with us.

RISKS, INCONVENIENCES, AND DISCOMFORTS: None.

ALTERNATIVES TO PARTICIPATION: You don't have to do this but if you don't we won't know how you felt about working with us.

CONFIDENTIALITY OF RECORDS: Nobody except your therapists will ever see your answers.

OTHER INFORMATION: Significant findings that occur during this study which might affect your decision to participate in the study will be discussed with you. Any significant findings

developed from this study will be available to you and may be obtained from the investigator. Your participation in this study may be terminated without your consent if conditions occur which might make your continued participation dangerous or detrimental to your health. The study itself may be terminated prior to your completing participation.

You are encouraged to ask any questions, at any time, that will help you to understand how this study will be performed and/or how it will affect you. If you have any questions, you may contact Dr. Owens at our clinic.

IF THERE IS ANY PORTION OF THIS EXPLANATION THAT YOU DO NOT UNDERSTAND, ASK THE INVESTIGATOR BEFORE AGREEING TO PARTICIPATE IN THIS STUDY. You will be given a copy of this consent document for your records.

SIGNATURE OF VOLUNTEER	DATE	
PERMANENT ADDRESS OF VOLUNTEER	<p style="text-align: right;">TYPED NAME OF WITNESS</p> <p style="text-align: right;">SIGNATURE OF WITNESS DATE SIGNED</p>	

Sample 4

Protocol using non-human animals

A word about formats for animal use protocols: Every institution and groups seems to have different, exceedingly odd formats for their animals use protocols. This is to some extent a response to problems with animal rights groups attempting to use the freedom of information act to identify and verbally harass investigators. The odd formats put the investigators' names and identifying information in places which are removed easily prior to being sent in compliance with the act. The protocols also get somewhat repetitious because different regulations and committees require the same information phrased or set out in different ways. Thus, do not use the following format as a guide when you write an animal use protocol. You will find someplace where all of the following information has to go.

About the protocol itself:

The following fictitious study is based on an actual study accepted for performance.

As you read this protocol, note the incredible amount of repetition required by various laws. Also note that the actual background, training, and experience of the investigators does not appear anywhere in the protocol itself.

Please also note that therapeutic touch is far better demonstrated than indicated in this protocol and that the references were made-up. However, the part about using non-invasive techniques to evaluate wound healing is accurate.

LABORATORY ANIMAL CLINICAL INVESTIGATION PROJECT
Sample University

PROTOCOL TITLE: Effects of therapeutic touch on wound healing and the development of a non-invasive assessment system for wound healing.

PRINCIPAL INVESTIGATOR: Julie Cutter, MD, PhD
Associate Professor of Surgery,
Surgical Research Service,

Sample University

CO-INVESTIGATORS: Lucy Lightlife, RN, PhD
Surgical Nursing Service
Sample University

Mark Seeview, MD
Assistant Chief,
Radiology Service
Sample University

Ann Ratbite, D.V.M.
Chief, Veterinary Pathology
Lickspite Consultants

Marvin Twister, PhD
Biomechanics Consultant

I. NON-TECHNICAL SYNOPSIS:

Therapeutic Touch is used to speed recovery from numerous diseases and to decrease pain and discomfort. Therapeutic touch has frequently been claimed to speed the healing of surgical incisions but this claim has never been objectively tested. One reason it has not been tested is that there is tremendous variability in the way and rate wounds heal depending on the incision and the condition of the person incised. A second reason is that there is no non-invasive way to tell how completely a wound has healed. There is little interest in simple closing of the outer layer of skin over a wound as the outer layer can look firm and healthy without any healing of the major tissues underneath. None of the usual ultrasound or x-ray methods have been well correlated with actual strength and amount of wound healing.

The study will consist of a pilot using two rats followed by a placebo controlled design. The pilot will consist of having the therapeutic touch expert familiarize herself with several rats during which she will get a "feel" for their "fields." The rats will then be anesthetized and slits made in their back skins and sewn up. The healer will spend several days accustomizing herself to differences in the way the rats feel to her after the anesthesia and incisions as well as in trying to help their healing. The second portion of the study will not take place unless the healer feels that she can help the rats.

The placebo controlled design will use two groups of 10 male rats each. The procedure will involve making full thickness incisions through the skin of rats' backs, sewing them up, and then waiting periods of one through five days before cutting out the incision area for both histological and tensile strength testing. Two rats from each group will be tested for wound healing for 5 consecutive days. The rats will be exposed to actual or placebo therapeutic touch for a minimum of four quarter hour periods every day starting the day of surgery and continuing through the day the incisions are removed. The rats will receive analgesics as needed and should not be in any

significant discomfort during the study. The rats will be sewn up after the incisions are removed and returned to their colony to live out their natural lives. Male rats are used to avoid the complications of and variability induced by pregnancy on wound healing.

The purpose is to see if exposure to therapeutic touch potentiates healing. Strength of the wound will be objectively tested by removing the wound site from the rat and recording the force required to pull the edges of the wound apart. Healing will also be assessed through standard histological techniques.

The objective measures of healing will be correlated with the results of non-invasive ultrasound measurements made just before the wound sites are removed. It will then be possible to determine the strength and □completeness□ of a healing a full thickness skin wound in humans simply by looking at an ultrasound. Without an animal study to calibrate non-invasive methodologies, tiny biopsies in full thickness wounds would have to be made in human subjects.

II. Background

A. Use of therapeutic touch for wound healing: Therapeutic touch for wound healing frequently consists of holding one's hands over a patient (no actual physical contact is usually made), sensing the wound, and channeling one's body energies to the patient to speed healing. This technique should not be confused with the very well documented increase in immune system functioning and increased recovery from innumerable diseases provided by human touch, caring, and holding (e.g. see review by Grayson of deaths in orphanages, 1998). There is no demonstrated mechanism of action for potentiation of wound healing through therapeutic approach and all attempts to show an effect on the body's magnetic field have failed (e.g. Gowen, 1961, Marksense 1993, Lasker 1998). While there is evidence that therapeutic touch can help some patients recover from some problems, its use for wound healing is based on clinical case reports, poorly designed group studies, and oft repeated stories of clinical success which spread like wildfire through the nursing community. There have been no appropriately designed trials of trials of therapeutic touch for wound healing acceptable to any recognized clinical group (e.g. reviews by Malwowski 1997, Carnack 1998). Every attempt to replicate supposedly successful clinical studies or to perform group studies in which (a) the healer is carefully observed and (b) careful ratings of wounds before and after days to weeks of therapeutic touch are made by neutral observers have failed to show any support for therapeutic touch (same reviews as above). In spite of this total lack of supporting evidence, the approach has become progressively more well codified to the extent that many people, especially nurses, are now certified in its use for wound healing as well as the other, somewhat better substantiated, uses.

There have been no objective studies of this technique for wound healing. Such studies need to take place before its use becomes generally accepted simply through its widespread presence.

B. Non-invasive evaluation of wound healing: Numerous trials of high intensity / ultra high resolution ultrasound (Garble 1997, Wishing 1998) and various sorts of radiological procedures with and without contrast media (Zapp 1989, Flash 1998) have been unsuccessful in determining how well healed surgical incisions are because their results have never been correlated with the amount of healing or strength of full thickness incisions.

C. Literature Search:

1. Literature Source(s) Searched: DTIC, Tech Reports, Work Units, Medline, Agricola, FEDRIP

2. Date and Number of Search: DTIC Worksheet Submitted: Search completed 31 August, 1995. Additional search on Agricola and FEDRIP completed 31 October, 1995.

3. Key Words of Search: Therapeutic touch, wounds, healing, ultrasound.

4. Results of Search:

FEDRIP yielded 1 project that was applicable: Schafer, Mark E. (FY 1995). Project title: Ultrasound/collagen treatment of full thickness wounds.

AGRICOLA yielded no references that were applicable.

DTIC, Work Units Information Services, and Medline searches were also done.

III. OBJECTIVE\HYPOTHESIS:

A. Objectives:

1. Determine whether therapeutic touch accelerates the rate of healing of full thickness skin incisions which are sutured using standard techniques.

2. Correlate ultrasound measurements of healing skin wounds with accessed objective measurements (tensile strength and histology).

B. Hypotheses:

1. That therapeutic touch performed by a believing, certified expert, accelerates the rate of healing full thickness skin incisions which are sutured using standard techniques relative to the rate of healing when an untrained neutral person stands in the same position for the same period of time over similar animals.

2. That it will be possible to correlate ultrasound measurements of healing skin wounds with objective measurements (tensile strength and histology).

IV. MEDICAL RELEVANCE: Quicker wound healing should lead to less morbidity and mortality as well as reduced stays in the hospital and quicker return to work.

V. MATERIALS AND METHODS:

A. Experimental Design and General Procedures:

1. Experimental Design Overview:

a. A pilot study using two rats will be used for the purposes of (1) permitting the therapeutic touch expert and the neutral toucher to become familiar with rats

before and after incisions are made in their skin and (2) practicing and gaining precision using the ultrasound, Instron, and surgical techniques prior to the study.

b. The actual study will be a placebo controlled study with two groups of 10 rats. Two rats from each group will be tested for wound healing for 5 consecutive days. Full thickness surgical skin incisions made on the animal's dorsum will be repaired using skin staples and permitted to heal. Half of the rats will be exposed to therapeutic touch for four quarter hour periods spaced throughout each day including the day of surgery. The co-investigator performing therapeutic touch is certified in its use and has six years of experience using the technique with patients as part of her nursing practice. She believes that she has helped surgical incisions to heal faster and that she helped one of her dogs heal faster after a wound was sewn up. The placebo group will receive similar attention from an investigator with no knowledge of (or belief in her ability to perform) therapeutic touch to reduce the possibility that simple additional human contact increases the rate of healing. Strength of the wound will be objectively tested by excising the wound site from the rat and recording the force required to pull the edges of the wound apart using a standard tensile testing device. Healing will also be assessed through standard histological techniques. These objective measures of healing will be correlated with the results of non-invasive ultrasound measurements made just before the wound sites are removed. This might permit use of ultrasound to determine the extent of post-surgical wound healing.

2. Procedure for the full study: We propose to make eight horizontal full thickness incisions, each about two centimeters long, on the backs of 20 anesthetized adult rats, and close the wound sites appropriately. The rats will be randomized into two groups (I and II) by alternate assignment according to the order they are removed from their cages. After wound closure, the ten rats from group I will be exposed to therapeutic touch for four quarter hour periods spaced through the day while the rats from group II are being similarly exposed to non-therapeutic touch. On every day (for 5 days), following the day of the surgery, four rats (two rats from the therapeutic touch group and two rats from the control group) will be anesthetized again. For each rat, imaging will be performed, the wound areas will be cut out, and the skin will be reapproximated with steel staples. Steel staples reduce giant cell responses thereby limiting inflammation. When the rats wake up, they will be returned to their colony. Figure One shows the sub-group classification:

Figure One

Study Design	
Post surgery day	<u># animals per group</u> Therapeutic touch : Control
1	2 : 2

2	2 : 2
3	2 : 2
4	2 : 2
5	2 : 2

Starting from the most anterior incision, the eight areas of skin will be alternately assigned for either Instron testing or histological examination. Four areas of skin will be tested for strength using a tensile strength testing device (Instron). This test is performed by removing the steel staples, cutting off the superfluous edges, and then mounting the remaining skin from each side of the wound in chucks. The Instron pulls the chucks away from each other at a set speed and strength while measuring both elasticity and force required to separate the wound. The other four areas of skin will be evaluated histologically. By using 4 animals each day, the amount of healing on post surgery days 1, 2, 3, 4, and 5 can be evaluated. Two animals per group should be sufficient for determining variability in healing rates due to unanticipated problem such as disease. Sixteen sections (eight from each of two animals) should be sufficient for statistical comparisons given the expected variability.

Laboratory Animals Required and Justification:

1. Non animal Alternatives Considered: Yes. However, no models of "in vitro" wound healing have ever been developed which adequately compensate for the many variables introduced by a living animal. Furthermore, no models exist which can correlate non-invasive methodologies with histologic evaluations of wound healing. Computer models capable of imitating these healing interrelationships do not exist because fundamental information about them has never been established. We have no idea how to simulate therapeutic touch on a non-living system.

2. Animal Model and Species Justification: Due to skin differences, lower phylogenetic species than rats are inappropriate. Thus their usage is not applicable to humans. Comparability of the species chosen for study to the human: Rat skin has been shown to exhibit similar wound healing characteristics.

3. Laboratory Animals:

a. Genus & Species: Rats: Rattus norvegicus

b. Strain/Stock: Outbred: Sprague Dawley

**c. Source/Vendor: B & K Universal Inc.
CIRO# 94-21**

d. **Age:** Adult

e. **Weight:** 200 - 300 g

f. **Sex:** Male

4. **Total Number of Animals Required:** rats 22

5. **Refinement, Reduction, Replacement:**

a. **Refinement:** No animals are expected to die or be in chronic pain from participation in this study.

b. **Reduction:** The appropriate statistical tests can not be performed with less than four animals in a group.

c. **Replacement:** Replacement: Canines have been used in previous studies. We are using rats which are lower on the phylogenetic scale.

C. Technical Methods: Prior to surgical procedures each experimental animal's back will be shaved using electrical animal shears, in order to clear the site for the study incision. Once the animals are anesthetized, the dorsal portion will be prepped using a betadine and beta-brandt's solution. The dorsal portion will be draped in a sterile fashion. Using a 15-blade, eight 2cm full thickness, lateral incisions will be made. The incisions will be at a minimum of 1 cm apart. Incisions will then be repaired using skin staples. Special care will be taken to minimize handling of skin edges with pick-ups. A sterile dressing of xeroform gauze strips and kling-roll will be applied. Animals will then be returned to cage for recovery.

Post-operative days one through five, two rats from each group will have their Study incision excised using the following procedure: The dorsum of each animal will be prepped and draped in the above noted fashion following receiving anesthesia. using a 15-blade, the eight Study incisions will be excised in an elliptical fashion.

Following skin incision, the elliptical skin flap (containing the study incision) will be undermined using a small hemostat in an insert-and-spread fashion to free up any soft tissue attachments/adhesions. Hemostasis will be achieved using a battery electric cautery as indicated. The four elliptical wounds will be irrigated with sterile saline and the wounds will be closed using skin staples. A dressing consisting of xeroform gauze strips and kling-roll will be applied.

1. **Pain:** All animals will have multiple incisions. This will be accompanied by using general anesthesia and analgesics as required

a. **USDA (Form 18 3) Pain category:**

(1) **No Pain** 0 (#) 0% (Column C)

(2) **Alleviated Pain** 24 (#) 100%(Column D)

(3) **Unalleviated Pain or Distress** 0 (#) 0%
(Column E)

b. Pain Alleviation: During the post surgical recovery period the animal care staff will provide appropriate analgesics.

(1) **Anesthesia/Analgesia/Tranquilization:** Animals will be anesthetized with Ketamine 75mg/kg and xylazine 10mg/kg IM using 26 gauge needles. Post operative analgesics will be buprenorphine .05mg/kg given SQ by animal care staff.

(2) **Paralytics:** N.A.

c. Alternatives to Painful Procedures: None

(1) **Source(s) Searched:** Medline, Agricola, FEDRIP.

(2) **Date of Search:** Medline, March 30, 1998; Agricola, FEDRIP, March 31, 1998.

(3) **Key Words of Search:** pain, surgery, wounds, incisions, and therapeutic touch

(4) **Results of Search:** Medline produced 2 related sources. FEDRIP yielded one project that was related: Stein, Cristoph (FY 1995) Endogenous opiates in inflamed tissue and analgesia.

AGRICOLA yielded no related references.

d. Painful Procedure Justification: N.A.

2. Prolonged Restraint: N.A.

3. Surgery:

a. Procedure: Each experimental animal will have its' back shaved with electrical animal shears prior to surgical procedure to create the Study incision. Once the animals are anesthetized, the dorsal portion will be prepped using a betadine and beta-brandt's solution. The dorsal portion will be draped in a sterile fashion, and using a 15-blade, eight 2cm full thickness, lateral incisions will be made. The incisions will be at a minimum of 1 cm lateral of the midline and 1 cm apart (see Figure 3). Incisions will then be repaired using skin staples. Special care will be taken to minimize handling of skin edges with pick-ups. A sterile dressing of xeroform gauze strips and kling-roll will be applied. Animals will then be returned to cage for recovery.

b. Pre and Postoperative Provisions: Post-operative days one through five, two rats from each group will have their Study incision excised using the following procedure: The dorsum of each animal will be prepped and draped in the above noted fashion following receiving anesthesia. using a 15-blade, the eight Study incisions will be excised in an elliptical fashion (see Figure 4). Following skin incision, the elliptical skin flap (containing Study incision) will be undermined using a small hemostat in an insert-and-spread fashion to free

up any soft tissue attachments/adhesions. Hemostasis will be achieved using a battery electric cautery as indicated. The eight elliptical wounds will be irrigated with sterile saline and the wounds will be closed using skin staples. A dressing consisting of xeroform gauze strips and kling-roll will be applied.

c. **Location:** Animal Surgery Suite in Animal Housing Facility, Research Department.

d. **Multiple Survival Surgery Procedures:** N/A

4. **Animal Manipulations:**

a. **Injections:** Refer to Pain Alleviation, anesthesia/analgesia section.

b. **Biosamples:** Skin samples will be taken and evaluated using the Instron and histological techniques.

c. **Animal Identification:** Indelible ink will be used to identify rats.

d. **Behavioral Studies:** N.A.

e. **Other procedures:** Ultrasound evaluation

5. **Adjuvants:** N.A.

6. **Study Endpoint:** After the healed wounds are removed, the animals will recover and be returned intact and healthy to Animal Resources Service after participation in the study ends.

7. **Euthanasia:** None. After the healed wounds are removed, the animals will recover and be available for further use.

D. Veterinary Care:

1. **Husbandry Considerations:** Animals will be housed according to LASS SOP # 201 and in accordance with the Guide for the care and use of Laboratory animals.

a. **Study Room:** N.A.

b. **Special Husbandry Provisions:** N.A.

2. **Attending Veterinary Care:**

(1) All of the animals will have some discomfort after surgery. It will be minimized with analgesics.

(2) Methods for appropriate alleviation of pain and distress: Animals will be held off food for 12 hours and water withheld for 12 hours prior to surgery.

(3) Exceptions to the alleviation of pain: NONE

(I) Surgery will be performed by the principal and associate investigators under the guidance of the attending veterinarian and the Orthopedic Surgeon.

(ii) Post-operative location of animals: Animal Surgery Suite in Department of Clinical Investigations.

(iii) Post-operative care plan: Animals will be observed twice daily by animal care personnel. Evaluation for pain will be accomplished during these observations. At least one of these observations will be done by the attending veterinarian for the first five days post operatively. If the assessment of the animal's signs warrant analgesic, Buprenorphine hydrochloride will be administered (0.005 mg/kg, IM) on an as needed basis.

(iv) Length of time necessary to monitor animals following the surgical procedure: 2 hours.

(v) The surgical procedures are very minor in that the first only requires making four incisions in the skin and the second only requires removing the healed skin by making an ellipse around it and then closing it up. Thus, the second surgical procedure is theoretically a second surgery. However both are minor and should leave the animals in as good condition as when they began participation. The animals should be essentially normal within weeks of the surgery and available for other duties as assigned.

3. Enrichment Strategy: N.A.

E. Data Analysis: A repeated measures, two way analysis of variance will be performed on each outcome measure. The histological evaluations will be ratings so a non-parametric ANOVA will be utilized. The tensile strength tests are likely to be parametric so a parametric ANOVA will be utilized.

F. Investigator & Technician Qualifications/Training:

All investigators named in this study have demonstrated an understanding of the humane care and use of research animals. They have taken part in discussions of pertinent laws and regulations concerning the use of animals in biomedical research as required by Public Laws 89-544, 91-579, 94-279, and 99-198 (The Animal Welfare Act and Amendments). They are familiar with the concepts for the reduction or elimination of the use of animals and have concluded that there is a need for the use of animals in this study. They have been familiarized in the proper methods for minimizing and/or alleviating pain and discomfort in the animal species selected for study. They will either have an animal technician assigned to assist them who is an expert in the animal manipulation techniques required for this study, or have exhibited sufficient proficiency themselves to justify allowing them to work unassisted or without direct guidance from the laboratory animal veterinary staff. They have been advised on care and use policy at this institution and are aware of the established reporting mechanisms for observed deficiencies in animal care and treatment. They have used the information services of this medical center's medical library, as evidenced by information provided in Section II.

Veterinarian's Signature

VI. Biohazard/Safety: Normal Laboratory animal handling procedures will be followed to prevent any zoonotic disease transmission.

VII. ASSURANCES: The law specifically requires several written assurances from the P.I. It states that "research facilities will be held responsible if it is subsequently determined that an experiment is unnecessarily duplicative, and that a good faith review of available sources would have indicated as much."

(This section will state) As the Primary Investigator on this protocol I acknowledge my responsibilities and provide assurances for the following:

A. Animal Use: The animals authorized for use in this protocol will be used only in the activities and in the manner described herein, unless a deviation is specifically approved by the IACUC.

B. Duplication of Effort: I have made a reasonable, good faith effort to ensure that this protocol is not an unnecessary duplication of previous experiments.

C. Statistical Assurance: I assure that I have consulted with an individual who is qualified to evaluate the statistical design or strategy of this proposal, and that the "minimum number of animals needed for scientific validity are used."

D. Biohazard/Safety: I have taken into consideration, and I have made the proper coordinations regarding all applicable rules and regulations regarding radiation protection, biosafety, recombinant issues, etc., in the preparation of this protocol.

E. Training: I verify that the personnel performing the animal procedures/manipulations described in this protocol are technically competent and have been properly trained to ensure that no unnecessary pain or distress will be caused as a result of the procedures/manipulations.

F. Responsibility: The principal investigator will not permit an animal under his stewardship to suffer unduly. No animals should die or be in chronic pain from participation in this study. The animals will be observed daily by the animal care staff for signs of discomfort and infection. In the unlikely and unanticipated event that a participating animal develops an infection which is resistant to reasonable treatment for a reasonable length of time, the animal will be euthanized to avoid prolonged suffering. I acknowledge the inherent moral and administrative obligations associated with the performance of this animal use protocol, and I assure that all individuals associated with this project will demonstrate a concern for the health, comfort, welfare, and well-being of the research animals. Additionally, I pledge to conduct this study in the spirit of the fourth "R" which the Nation has embraced, namely, "Responsibility" for implementing animal use alternatives where feasible, and conducting humane and lawful research.

Principal Investigator's Signature

G. Painful Procedures: I am conducting biomedical experiments which may potentially cause more than momentary or slight pain or distress to animals that **WILL BE relieved or WILL NOT (circle one) be relieved with the use of anesthetics, analgesics and/or tranquilizers.** I have considered alternatives to such procedures; however, using the

methods and sources described in the protocol, I have determined that alternative procedures are not available to accomplish the objectives of the proposed experiment.

Principal Investigator's Signature

VIII. Enclosures:

- A. Resource Requirements (required):
- B. Impact Statements Histology and EM
- C. Literature Searches (required): DTIC, FEDRIP, MEDLINE, AGRICOLA, etc.
- D. Pain Scoring Guidelines:
- E. Resumes/CVs for all investigators and consultants:

ENCLOSURE A - RESOURCE REQUIREMENTS

Resources already available: One major equipment system required to perform the study, the tensile strength evaluation system (\$172,000), is currently on hand. All computer based statistical systems for storing, manipulating, and analyzing the data are on hand (SPSS-advanced and Statpack Gold plus on two Pentium computers.

- 1. **ANIMAL COSTS:** $\$15.60/\text{rat} \times 24 = \$374.40 + \$50$ (shipping & handling)= **\$424.40**
(Two extra rats are ordered in case of sickness before entry into the study.)
- 2. **PER DIEM COSTS (PER DAY):** \$0.20 rat/day
- 3. **ESTIMATED NUMBER OF DAYS TO BE HOUSED:** 24 animals for 30 days each.
- 4. **PROJECTED ANNUAL PER DIEM COST:** 24 rats X 30 days X .20/day= **\$144**
- 5. **CONSUMABLE SUPPLIES:** Staples, surgical supplies= **\$1,000**, Pathology and Histology **\$800** Total = \$1,800
- 6. **EQUIPMENT PURCHASE COSTS (Other than caging):** None.
- 7. **FACILITY MODIFICATION/CAGING COSTS:** None other than that supplied by Animal Resources Service.
- 8. **TOTAL COST OF ENTIRE STUDY:** Potential maximum of **\$2,368.40**
- 9. **GRANTS, GIFTS AND LOANED EQUIPMENT:** None at this time.

10. SOURCE OF GRANTS, GIFTS AND LOANED EQUIPMENT: N/A

ENCLOSURE B - IMPACT STATEMENTS

The Research Department is capable of providing the requested support for this protocol, as enumerated below. The support of this protocol will not adversely affect human patient health care delivery.

1. Chief of Research.

Signature of the Chief of Research

2. (List other service/departments providing support).

Chief, (Service/Department)

Bibliography: *Purposely left out as all references are fictitious.*

Sample 5

Poor paper

The following “paper” is modified from an actual manuscript submitted for publication. The names of the authors, subject of the paper, and much of the text have been changed to protect the authors. All of the references have been eliminated so readers can not determine the initial subject matter of the paper. I made up □mulehaulers syndrome□ to avoid compromising the introduction.

Biofeedback therapy for the management of mulehaulers syndrome

James Martin, M.D., Larry Owens, Ph.D., and Judy Knowhow, MA, BCIAC
Albright Surgical Center, Cutnose, Georgia

ABSTRACT

Mulehaulers syndrome is a neurovascular disorder affecting the upper extremities of the body. It is virtually always bilateral with patients reporting chronic, debilitating pain in the shoulders and arms. No treatments are known to be generally effective for this disorder and most patients undergo numerous unsuccessful medical, rehabilitative, and surgical procedures. Biofeedback is known to be effective for many painful disorders but has not be used for mulehaulers in the past.

This study investigated the effectiveness of biofeedback for the treatment of 63 sequential patients with mulehaulers syndrome. All patients underwent a through diagnostic examination and were given individualized treatment programs. The results show that patients benefited significantly from biofeedback. All of them reported that the intervention was helpful and most reported less pain and emotional distress with corresponding increased ability to function. Biofeedback is highly recommended as a primary intervention for mulehaulers syndrome.

Key words: pain, mulehaulers syndrome, biofeedback

INTRODUCTION

Mulehaulers syndrome is a neurovascular disorder affecting both upper extremities. It was originally reported among teamsters who had to pull mules forward by their head bridles and is now generally spread through the population of people who pull carts or hold heavy objects in their arms (1). It is virtually always bilateral with patients reporting chronic, debilitating pain in the shoulders and arms (1). The pain is usually stabbing in nature and both range of motion and

motor function are frequently impaired. No treatments are known to be generally effective for this disorder and most patients undergo numerous unsuccessful medical, rehabilitative, and surgical procedures (1,2). The disorder appears to be a distant cousin of thoracic outlet syndrome, whose actual etiology and underlying physiological problems are also unclear (3,4,5). It is likely that chronic mechanical pressure on blood vessels and nerves eventually causes the problem in those susceptible to it (4,6). In our hands, nerve conduction studies are frequently positive.

After diagnosis, our normal treatment consists of physical therapy, wrists and elbow splints, lifestyle modification, analgesics, nonsteroidal-anti-inflammatories, muscle relaxants, antidepressants, and nerve blocks. If these are not effective within ten weeks, surgical decompression of the nerves and blood vessels is attempted for selected patients. If all of this fails, the patients are referred for our biofeedback pain management program.

Biofeedback is known to be effective for many painful disorders (3,4,5) but has not been used for mulehaulers in the past. This study was intended to evaluate the contribution of biofeedback to amelioration of the chronic pain and disability among patients diagnosed with mulehaulers syndrome who failed conservative and surgical intervention.

METHODS

All of the participating patients completed our clinic's biofeedback pain management program (a minimum of 15 sessions). The program includes individual training in sEMG and hand temperature biofeedback, cognitive and behavioral techniques, patient education, positive reinforcement, progressive muscle relaxation training, body scanning, autogenic training, visual imagery training, diaphragmatic breathing. The specific training given to each patient is tailored to his or her needs.

Biofeedback was given using an A&L 663 (Avicron Labs, NJ) with one channel for temperature training and either one or two for sEMG training. Visual feedback in the form of graphs was provided on color monitors and therapists gave verbal feedback at the end of each session.

After completion of the program, each patient filled out a post-treatment questionnaire which consisted of ten Likert scaled questions on (1) the pleasantness of the sessions, (2) the helpfulness of the therapists, (3) the helpfulness of the biofeedback, (4) decrease in discomfort, (5) the ability to deal with discomfort, (6) ability / stamina to engage in more activities, (7) enjoyment of participation in physical activities, (8) feelings of unhappiness, (9) feelings of nervousness, and (10) amount of perceived control over life.

RESULTS

The questionnaire was given to 63 sequential patients who completed the above program over a six year period. They were treated by any of seven therapists who worked in the clinic during that period of time using the above protocol. The subjects completed an average of 21.4317 sessions (SD= 31.3971). Sixteen of the subjects were males (mean age 38.5971 +/- 15.49) and

47 were females (mean age 23.2939 +/- 10.38).

Overall, the results show that patients benefited significantly from biofeedback. All of them reported that the intervention was helpful and most reported less pain and emotional distress with corresponding increased ability to function. The means, standard deviations, etc. of the results are detailed and compiled in the table. The overall results for each question were as follows:

- (1) Pleasantness of the sessions: 96% of the patients rated the sessions as pleasant.
- (2) Helpfulness of the therapists: 93% rated the therapists as being helpful.
- (3) Helpfulness of the biofeedback: 91% rated the biofeedback as being helpful.
- (4) Decrease in discomfort: 71% rated their discomfort as being less than when they started the program.
- (5) Ability to deal with discomfort: 83% rated their ability to deal with their discomfort as being greater.
- (6) Ability / stamina to engage in more activities: 66% rated an increase in ability and stamina.
- (7) Enjoyment of participation in physical activities: 64 % rated an increase in enjoyment.
- (8) Feelings of unhappiness: An amazing 86% of the patients rated themselves as being less unhappy after participation in our program.
- (9) Feelings of nervousness: 81% rated themselves as less nervous.
- (10) Amount of perceived control over life: 82 % felt an increase in control over their lives.

Two percent of the participants apparently did not feel they had gained anything from participation since their ratings on the Lacerate scale were universally negative.

DISCUSSION

It is clear from these results that biofeedback significantly relieves the pain and symptoms of virtually all patients with mulehaulers syndrome. In view of these clear results, we strongly recommend that biofeedback be provided to all patients with this disorder.

Our clinic has been charging patients for this therapy through out its existence. We began billing insurance carries for the therapy on the basis of these findings. We were surprised when numerous insurance companies refused to reimburse us for this therapy even when supplied with copies of our study. We feel that their decisions are prejudicial and unjust so are appealing through the various company's hierarchies.

We recently began using neurofeedback in lieu of sEMG and temperature biofeedback with even more spectacular results. Nearly all of the subjects report enormous decreases in nervousness and increases in energy for performing activities. We plan to issue a follow-up report based on our findings with the first ten patients treated under the new regime as soon as they have completed therapy.

(REFERENCES PURPOSELY LEFT OUT)

Table 1

Scale: 1- 3 = negative; 4 = neutral or no change; 5-7 = positive

Rated Item	% rated positive	mean rating	SD of rating
(1) Pleasantness of the sessions	96%	6.12	2.1
(2) Helpfulness of the therapists	93%	6.94	0.326
(3) Helpfulness of the biofeedback	91%	6.1	3.7
(4) Decrease in discomfort	71%	5.1	4.87
(5) Ability to deal with discomfort	83%	5.916	0.9
(6) Ability / stamina to engage in more activities	66%	4.865	1.61
(7) Enjoyment of participation in physical activities	64 %	4.37	1.3951
(8) Feelings of unhappiness	86%	5.28	1.17
(9) Feelings of nervousness	81%	5.1	2.3
(10) Amount of perceived control over life	82 %	5.961	2.153

Sample 6

Paper based on the pilot study in sample 1

Citation:

Interplanetary Journal of Scientifically Substantiated Instant Cures, 16: 236 - 247, 1998.

(Please note that the following is a very modified version of an article which actually appeared in a real, highly reputable, journal. This work was performed by employees of the US Government so can not be copywrited so it's use here is not in contravention of the law. I apologize for any appoleptic attacks suffered by the actual editor if he happens to come across this modification.)

Treatment of migraine headaches using pulsing electromagnetic fields***

*James E. Expert, Ph.D., *Linda A. Example, BA, and **Lucey Nervesplit , MD

*Sample Institutes, Hometown, Washington
and **Neurological Associates, Inc. Saffron, WA ,

*** Most of the data were presented at the August 1997 World Congress on Instant Cures held in Marlboro, Newschotlund

Running head: Pulsing magnetic fields for migraines

SYNOPSIS

In an open study, eleven chronic migraneurs (2 males and 9 females) were exposed to pulsing electromagnetic fields (PEMFs) over their inner thighs. Subjects kept a one month week headache log before and after two to three weeks of exposure to PEMF for one hour per day, five days per week. Number of headaches per week decreased from 4.03 during the baseline period to 0.43 during the initial one month follow-up period and to 0.14 during the extended follow-up which averaged 8.1 months. Large controlled studies should be performed to determine whether this intervention is actually effective.

INTRODUCTION

Most adults in the United States have at least occasional headaches. Headache is now the leading medical cause of lost days of work and costs the U.S. many billions per year to treat. The Nuprin Pain Report (1987) found that 157 million work days per year were lost due to this problem alone. Numerous surveys of the general population have indicated that about 65 percent of males and 78 percent of females reported having had at least one headache within the past year (1,2). About half of the men and 65 percent of women report having at least one headache per month. About 30 percent of males and 44 percent of females reported that these were severe with about 15 percent having headaches severe enough to affect daily activities. Among young adults, severe headaches lasted about six hours for males and eight for females with eight percent of males and fourteen percent of females losing a day or more of work per month due to headaches. About six percent of people attending civilian general medical practices request treatment for headaches as their primary reason for coming. About seven percent of males and seventeen percent of females reporting headaches requested treatment within the previous year (3).

Pulsing electromagnetic fields (PEMF) have been in use as therapeutic modalities for at least forty years. One of the well recognized, standard uses of PEMF generators is in enhancing the rate of healing of non-union fractures (4). The investigators were treating a patient for such a knee condition when the patient mentioned that she had a long history of weekly migraine headaches with auras (classic migraines) which had stopped shortly after PEMF treatment began. She reported that the visual auras had continued to occur as usual but the subsequent headache did not follow. Her headaches did not return for months after the end of treatment. This effect led us to wonder how exposure to PEMFs at the upper leg could have an effect on headaches presumably centered in the head.

The PEMF units used in our clinic (Diapulse model D103; Diapulse INC. of New York) are set to produce pulsed high-frequency, high peak power electromagnetic energy at a frequency of 27.12 MHZ in 65 microsecond bursts occurring in 600 pulse per second sequences at 975 peak watts. This is sufficient power to light a sixty watt bulb placed into the field. The field extends about 12 cm from the unit's head in a conical pattern. The unit's head is placed just above the area to be exposed and turned on for a set amount of time. Various units from other manufacturers differ slightly in a variety of ways such as the exact shape of the wave, rise and fall times, and power. The device looks like a floor mounted hair drier from the 1950s, has a relatively loud fan, a ticking timer, and sufficient knobs, lights, meter, etc. to be quite impressive. This impression has to be considered when attempting to differentiate actual from placebo effects.

Exposure to pulsing electromagnetic fields of the type described above appear to result in at least a temporary increase in peripheral blood flow. For example, Erdman (5) recorded peripheral blood flow from twenty normal subjects using both a temperature probe and volumetric measurements while they were being exposed to PEMF generated fields. He found a high correlation between the amount of energy produced by the device and peripheral blood flow with increases beginning within about eight minutes and plateauing by 35 minutes. Pulse rate and rectal temperatures did not change. This relationship has been confirmed in basic studies of blood flow in rabbit ears (6). Ross (7) recently reviewed the basic science and animal studies as well as some of the clinical studies showing the effectiveness of PEMF generators in increasing blood flow and wound healing. Cameron (8) demonstrated increased rates of healing in

experimentally induced wounds in dogs. Goldin et al (9) found similar results among humans in a double blind study using changes in fibroblast concentration, fibrin fibers, and collagen in the wound sites and in swelling. The increased rate of wound healing was ascribed to increased blood flow. Thus, it is very likely that exposure to this device does result in increased peripheral blood flow at least while exposure is in progress.

Freedman (10) has reviewed the evidence for temperature biofeedback's effect on peripheral blood flow. Numerous double blind studies with five to fifteen year follow-ups have demonstrated that training migraine headache patients to increase peripheral blood flow, through such techniques as temperature biofeedback from the finger, results in sustained decreases in all aspects of headache activity among a large percentage of people who successfully learn the techniques (11). Thus, whatever other mechanisms come into play, a technique which is aimed solely at increasing peripheral blood flow, frequently results in decreased headache activity when peripheral blood flow is successfully increased.

Synthesis of this background material led us to believe it was possible that application of a pulsing electromagnetic field directed only to the thigh, with its significant vascular supply, could produce sufficient increases in peripheral blood flow to effect headache activity for as long as the effect on blood flow continued. The following explorations were attempts to begin testing that hypothesis.

METHODS

Eleven (9 females, 2 males; mean age of 46.4 +/-14.9 years with a range of 20 - 73 years) having multi-year histories of headaches (mean of 17.9 years +/- 9.9 with a range of 2 - 40 years) were recruited from among the patients at a large military medical center. All of the subjects met the normally accepted criteria for migraine headaches with or without aura (common or classic migraines) set out by the 1982 Ad Hoc Committee for Headache Classification (11). Several of the subjects had mixed headaches because they reported headaches meeting both the above criteria for migraine headaches and those for tension headaches. After listening to an explanation of the study, each was given a consent form approved by the center's IRB to read and sign. Demographic and diagnostic data for the participants in the open study are presented in table one.

Each of the eleven patients kept a one month diary of headache activity before and after being exposed to PEMFs. Several studies reviewed by Blanchard et. al (12) indicate that one month week should be a sufficient baseline to determine the actual baseline level of activity in this population. The structure of the open study was essentially an "ABA" design with non-exposure periods surrounding a single exposure period. The subjects' prophylactic medications were stopped abruptly during the weekend between the end of the two week initial log and the start of the treatment period. Thus, any rebound effect should have come by the end of the treatment period and have been observed during the last week of treatment or the two initial post-treatment weeks. During the exposure period, patients were exposed to between two and three weeks of PEMF treatments for one hour per day, five days per week at a power of 975 watts with 600 pulses per second. The PEMF was applied to the medial thigh. Headache activity during treatment was reported to the therapist every day. The PEMF unit utilized in the study was described in the introduction. Additional follow-ups beyond the one month post-treatment log period were collected telephonically for all of the subjects.

The data were analyzed using repeated measures analyses of variance to determine whether there was an overall difference in headache activity between pre-treatment, post-treatment, and follow-up periods. Paired t tests were used to determine whether there were differences between any two periods when the analysis of variance was significant. The data met the entrance criteria for the tests.

RESULTS

The results for each individual are presented in Table 1. The average number of headaches per week decreased from 4.03 (+/- 2.02) during the two week pre-treatment baseline period to 0.43 (+/- 0.36) during the two week post-exposure period (statistically different at $p=0.001$; paired " t "= 5.998 with 10 DF). Follow-up data for the subjects are reported in Table 1. Follow-ups ranged from 1 week to 14 months with a mean of 8.1 months and a standard deviation of 3.09 months. The average number of headaches continued to decrease during the long term follow-up period to an average of 0.14 (+/- 0.08) per week (statistically different at $p=0.001$; paired " t "= 5.77 with 9 DF). A one way, repeated measures analysis of variance indicates that there was an overall difference between the periods ($p = 0.0001$, $F= 31.21$).

Table one about here

DISCUSSION AND CONCLUSION

Most of the patients with long histories of vascular migraine headaches not initiated by other problems who were exposed to adequate pulsing electromagnetic fields showed a dramatic decrease in headache activity during the weeks of exposure and for months afterwards. It is possible that exposure to PEMF did cause sufficient increase in peripheral blood flow to effect headache activity. This increase would have initiated some chain of psychophysiological events which actually caused the improvement. Of course, it is quite possible that exposure to the fields induced a currently unrecognized physiological effect, having nothing to do with blood flow, which somehow resulted in decreased headache activity. Interestingly, none of the patients who had aura preceding onset of their headaches reported a change in aura activity. The usual headache simply did not follow the aura.

The literature reviewed in the introduction shows considerable evidence that the particular PEMF generator used in this study has some ability to increase peripheral blood flow. The same body of evidence does not seem to exist for the weaker, battery powered units, magnetic field generators which do not pulse, or permanent magnets that people strap to themselves for a variety of reasons. If the working hypothesis (that increased peripheral blood flow has resulted in decreased headache activity) is correct, then devices not capable of increasing peripheral blood flow to a similar extent may not be effective. On the other hand, any device which can safely increase peripheral blood flow sufficiently, should be equally efficacious. PEMF units may produce increases in blood flow simply by heating the underlying tissues (13) rather than through any esoteric effects of their fields so it is possible that diathermy units could also produce these effects.

In spite of the impressive nature of the PEMF generator, the investigators do not feel that the major effects were due to placebo responses because (a) the participants each had multi-year histories of unsuccessful treatments with numerous highly touted therapeutic approaches, (b) the change in headache activity was much greater than would be anticipated due to a placebo response, □ the decrease in headache activity has been maintained longer than the six months or so anticipated for a placebo effect, and (d) the rebound effect from stopping the prophylactic medicines should have hit sometime near the end of “treatment” so would have overwhelmed any placebo effect.

Because of the powerful effects demonstrated in this trial, it is worth performing (a) larger double-blind, controlled studies to determine whether this intervention is actually effective and (b) longitudinal studies to determine whether inexpensive, wearable magnetic devices or other means of increasing peripheral blood flow can be used to keep the headaches from returning.

ACKNOWLEDGMENT

This study was entirely supported by the Sample Institutes.

REFERENCES

1. Linet M, Stewart W, Celentano D, Ziegler D, Sprecher M. An epidemiologic study of headache among young adults. JAMA. 1989; 261: 2211 - 2216.
2. Pascual J, Polo J, Berciano J. Serious Migraine: A study of some epidemiological aspects. Headache. 1990; 30: 481 - 484.
3. Blanchard E, Andrasik F. Management of chronic headaches. New York: Pergamon Press, 1985.
4. O'Connor M, Bentall R, Monahan J. Emerging Electromagnetic Medicine. New York: Springer-Verlag, 1990.
5. Erdman W. Peripheral blood flow measurements during application of pulsed high frequency currents. Am J of Orthopedics. 1960; 2: 196-197.
6. Fenn J. Effect of PEMFs (Diapulse) on experimental hematomas. Canadian Medical Association Journal 1969; 100: 251- 254.
7. Ross J. Biological effects of PEMFs using Diapulse. Chapter in : (M. O'Connor, R. Bentall, and J. Monahan, Eds) Emerging Electromagnetic Medicine. Pages 269 - 281. New York: Springier-Verlag, 1990.
8. Goldin J, Broadbent N, Nancarrow J, Marshall T. The effects of Diapulse on the healing of wounds: a double-blind, randomized controlled trial in man. Brit J of Plastic Surg 1981; 34: 267

- 270.

9. Cameron B: Experimental acceleration of wound healing. Am J of Orthoped 1961; 3: 336 - 343.

10. Freedman R. Physiological mechanisms of temperature biofeedback. Biofeedback and Self-Regulation 1991; 16(2): 95 - 115.

11. Tollison D, Kunkel R. Headache: Diagnosis and treatment. Baltimore, Williams & Wilkins, 1993.

12. Blanchard E, Hillhouse J, Appelbaum K, Jaccard J. What is an adequate length of baseline in research and clinical practice with chronic headache? Biofeedback and Self-Regulation 1987; 12(4): 323-329.

13. Wildervanck A, Wakim K, Herrick J, Krusen F. Certain experimental observations on a pulsed diathermy machine. Archives of PM&R 1959, 40: 45 - 55.

Table 1:
Open study exposing headache patients to pulsed electromagnetic fields

Migraine with Non-Visual Precursors or Auras			
Demographics (sex, age in years, number of years of headaches, cause of headaches)	One month, pre-treatment log of headaches per week	One month, post-treatment log of headaches per week	Duration of Follow-Up and number of headaches per week
1. Female 22 yrs old 12 yr history	7	0.25	9 month FU 0.13 HA/week
2. Male 21 yrs old 2 yr history	3	0.5	9 month FU 0.08 HA/week
3. Female 41 yrs old 19 yr history	2	0.5	9 month FU 0.13 HA/week
4. Female 43 yrs old	2	0	14 month FU 0.25 HA/week

21 yr history			
5. Female 38 yrs old 15 yr history	4	0.5	10 month FU 0.06 HA/week
6. Female 71 yrs old 40 yr history	6	1	9 month FU 0.08 HA/week
7. Female 73 yrs old 18 yr history	7	0	8 month FU 0.25 HA/week
8. Female 64 yrs old 6 yr history	4.3	1	1 week FU

Migraine with visual Auras

1. Female 20 yrs old 2 yr history	2	0.5	8 month FU 0.08 HA/week
2. Female 39 yrs old 11 yr history	2	0	7 month FU 0.25 HA/week
3. Male 38 yrs old 25 yr history	5.3	0.5	5 month FU 0.13 HA/week

Sample 7

Grant application based on the pilot study in sample 1

Application to:

The Philanthropic Society for the Encouragement of Deranged Experiments on Humans

The application **precisely** follows the outline and requirements supplied by the granting society.

(Note: This is a modified version of an application submitted to a very reputable foundation which was kind enough to fund the proposed project and, thus, made the work proposed in sample two possible. They liked the results of the study based on sample two sufficiently to fund further work in the area.)

Grant submission to the _____

Title: Headache treatment with pulsing electromagnetic fields

Date request submitted: 23 June, 1996

Request by: James E. Expert, PhD
Hometown, WA (206) 819 - 6423
Staff, Sample Institute

Lucey Nervesplit, MD
Neurological Associates, Inc.
Saffron, WA (360) 149 - 3948

James R. Frank, DO
Medical Director
Neurological Associates, Inc.
Saffron, WA (360) 149 - 3948

Request from: The Sample Institute, Hometown WA

Name and contact information for responsible business officer: _____

Total funds requested from the granting agency: \$10,000.

Duration of the project: 1 year.

1. SUMMARY OF REQUEST: The standard use of pulsing electromagnetic fields is treatment of non-union fractures. A patient being treated for non-unions reported that her migraines had stopped during treatment. An open pilot study showed that headache activity among eleven patients with migraines decreased from an average of 4.4 headaches per week to 0.6 per week for several to ten months after, two to three weeks of daily exposure of their thighs to pulsing electromagnetic fields (975 watts, 27.12 MHZ fields with 65 microsecond bursts pulsing 600 times per second). The proposed study will determine whether the results of the pilot were due to placebo effects. Participants will have at least a four year history of having migraine headaches with aura at least once per week, be of either sex, and between the ages of 18 and 70. Subjects will keep a daily log of the frequency and intensity of headaches as well as medication use for one month. They will then be randomized into actual or placebo pulsing electromagnetic field exposure groups and be exposed to pulsing electromagnetic fields (real or placebo) on alternating thighs for one hour per day, five days per week for two weeks. A different therapist will gather the data than will expose the patients to maintain the double blind condition. Neither the therapist gathering headache data nor the patient will know which group they are in as patients can not sense the fields and both real and placebo machines sound and look alike. Next,

patients will keep a one month "follow-up" log. A power analysis of the pilot data indicates that twenty subjects will be required per group. The proposed study was attempted without funding but only one therapist was available. It took the therapist less than two weeks to figure out which generator was the placebo because of differences in patient's reports about headache activity. Thus, the study must be performed with two therapists. The proposed design will cost approximately \$63,600 (including \$50,000 for the two generators) to perform. The institution can cover all but the \$10,000 required to pay the technicians.

2. SPECIFIC AIMS:

a. Relationship to the laboratory's overall program: The work proposed brings together our long term exploration of headache mechanisms and interventions (e.g. _____ 1982, 1991) with our relatively recent, grant supported, efforts to evaluate the effectiveness of pulsing electromagnetic fields for treatment of pain related disorders including tibial stress fractures (_____ and _____ 1991), pelvic stress fractures (_____ & _____ 1994), ankle sprains (_____ and _____ 1996), and post-surgical wound healing (_____ and _____ 1993). The program forms an integral part of our attempt to assess to the usefulness of alternative medical approaches to pain problems (e.g. _____ et al's 1994 NIH grant to evaluate the effectiveness of biofeedback for orofacial and low back pain and Osgrand and _____'s 1991 grant from the Veterans Administration to establish the reliability of psychophysiologic recordings of pain patients).

b. Relationship to the laboratory's work on effects of pulsing electromagnetic fields on headache activity: The proposed project grew out of our observation that patients being treated with pulsing electromagnetic fields for fractures reported a cessation of headaches during treatment. Our first study exploring this effect was an open trial in which people with migraine headaches kept multi-week logs of headache activity were exposed to pulsing electromagnetic fields. This exposure produced a significant decrease in headache activity (described in the preliminary studies section below) so the current, double blind study, is being proposed to determine whether the effects are due to placebos. If this study shows that the effects are real and sustained, further work will be proposed to (1) determine the optimal rate at which follow-up treatments have to be given to sustain any effects, (2) the duration of any effects observed, (3) whether exposure to pulsing electromagnetic fields also effects migraine headaches without aura and tension headaches, and (4) determine whether inexpensive, □home use□ devices will have a similar effect.

c. Objective: To determine whether migraine headache (with aura) activity changes from baseline during pulsing electromagnetic field and placebo exposure periods.

b. Hypothesis: That exposure to pulsing electromagnetic fields for two weeks will result in decreased migraine headache activity relative to decreases produced by exposure to a placebo pulsing electromagnetic field generator based on differences in headache activity determined from one month pre and post-exposure logs.

3. BACKGROUND AND SIGNIFICANCE:

a. Background:

(1) Incidence, impact, and request for treatment of migraine and tension headaches in the general population: Most adults in the United States have at least occasional headaches. Headache is now the leading medical cause of lost days of work and costs the U.S. many billions per year to treat. The Nuprin Pain Report (1987) found that 157 million work days per year were lost due to this problem alone. Numerous surveys of the general population have indicated that about 65 percent of males and 78 percent of females reported having had at least one headache within the past year (Linnet et al 1989, Pascual et al 1990). About half of the men and 65 percent of women report having at least one headache per month. About 30 percent of males and 44 percent of females reported that these were severe with about 15 percent having headaches severe enough to affect daily activities. Stewart et al's (1992) massive survey of 15,000 households also found this range of incidence. The data hold for young adults of typical military age. Among young adults, severe headaches lasted about six hours for males and eight for females with eight percent of males and fourteen percent of females losing a day or more of work per month due to headaches. About six percent of people attending civilian general medical practices request treatment for headaches as their primary reason for coming. About seven percent of males and seventeen percent of females reporting headaches requested treatment within the last year (Blanchard and Andrasik 1985).

(2) Availability and similarity of the subjects in the proposed setting to the general population: This material is presented to demonstrate that the population the investigators plan to work with is similar to the general population and is available to participate in headache studies.

(a) A pilot study was conducted by several of the investigators in 1993 to 1994 to investigate the prevalence and impact of headaches among people in waiting rooms at this facility. The 78 participants were not waiting for treatment for pain problems. The surveyors reported that virtually everyone asked to fill out a survey did so. Of the 78 responses, 83% said they had headaches with 57% taking over the counter medications and 17% taking prescription medications for their pain. 21% reported that they had to leave work due to headaches an average of 8.7 (SD 9.9) hours per month. Sufficient surveys were returned so that we can be 95% certain (+/- 10%) that these data represent the general population surveyed.

(b) A survey by several of the investigators was conducted of all 450 healthy adult women who came to this facility for check-ups last March. Virtually all of the surveys were returned filled out. Only one question dealt with headaches. It asked if they had headaches and, if so, the number of days per month their work was effected by the headache and the average severity on a scale of zero (no pain) to ten (so much pain they would be entirely incapacitated). One hundred and ninety (42%) of them indicated that they had headaches. The average severity was 5.7 (standard deviation = 2.4) and the average number of days per month work was effected equaled 6.2 (standard deviation = 5.9). No similar surveys were given to males we were investigating health problems of females as part of a Women's health initiative funding program.

A review of the use of the ER at this facility for just the first six days of April, 1996 (conducted by the principal investigator) showed that 16 patients with the chief complaint of headache were seen during that time. An average of 2.5 hours of patient contact time were required to care for these patients. A similar review for 7 - 17 April identified 45 patients.

(d) It should be noted that several of our headache research studies (_____ and _____ 1991, _____ et al 1991, 1992, 1993) which utilized volunteers from the community served by this facility who knew that they would never be identified and that their participation would never be noted in their medical records showed male to female ratios of 11:30 and 5:9. The ratio for a low back pain treatment study was 15:4. Thus, our population of males do get headaches but may not request treatment for them from normal medical channels. This is crucial because many practitioners frequently have the impression that severe headache is mainly a problem among women.

(2) Use of pulsing electromagnetic fields:

(a) Typical pulsing electromagnetic field generators and their fields: This technology has been in use since the 1950s. Units of the type we propose to use produce pulsed high-frequency, high peak power electromagnetic energy at a frequency of 27.12 MHZ in 65 microsecond bursts occurring in sequences ranging between 80 and 600 pulses per second. Wattage ranges from 293 to 975 peak watts for some units and less for others. Both pulses per second and wattage can be set in any of six steps. The field extends about 12 cm from the unit's head in a conical pattern. The unit's head is placed just above the area to be exposed and turned on for a set amount of time. The units look like floor mounted hair driers from the 1950s. They have a relatively loud faint, a ticking timer, and sufficient knobs, lights, meter, etc. so are quite impressive to be around. This impression has to be considered when attempting to differentiate actual from placebo effects. A typical generator is illustrated in Figure One. Various units differ slightly in a variety of ways such as the exact shape of the wave, rise and fall times, and power output. There is no actual evidence that any of these differences have any clinical importance. Most of the following studies were performed with the model we propose to use (Model D103, Diapulse INC. of New York), so the results may not apply to other devices. The FDA permits marketing of the pulsing electromagnetic field device to be utilized in this study.

Figure One:

Typical Pulsing Electromagnetic Field Generator

Photograph of Diapulse model D103.

(Photo same as one used in earlier example so not shown.)

(b) Pulsing electromagnetic fields and blood flow: Erdman (1980) recorded peripheral blood flow from twenty normal subjects using both a temperature probe and volumetric

measurements while they were being exposed to pulsing electromagnetic field generated fields. He found a high correlation between the amount of energy produced by the device and peripheral blood flow with increases beginning within about eight minutes and plateauing by 35 minutes. Pulse rate and rectal temperatures did not change. This relationship has been confirmed in basic studies of blood flow in rabbit ears (Fenn 1969). Ross (1990) recently reviewed the basic science and animal studies as well as some of the clinical studies showing the effectiveness of pulsing electromagnetic field generators in increasing blood flow and wound healing. Cameron (1961) demonstrated increased rates of healing in experimentally induced wounds in dogs. Goldin et al (1981) found similar results among humans in a double blind study using changes in fibroblast concentration, fibrin fibers, and collagen in the wound sites and in swelling. The recent status of the emerging field of electromagnetic medicine has been reviewed in a book by O'Connor et al (1990).

Clinical uses of pulsing electromagnetic field generators: These devices have recently been used very successfully by the Army in a study on treatment of grade I and II ankle sprains (Pennington et al 1993). Pennington's article reviews the safety of the technique and its usefulness for speeding recovery and reducing swelling. Pennington et al (1993) found that edema from grade one and two ankle sprains was reduced faster among 25 soldiers than among 25 control soldiers. We (_____ and Robson, 1996, in preparation for submission) have just completed an extension of this study. Wilson (1972) performed a placebo controlled study in which forty patients with inversion ankle injuries were paired for sex, age, weight and degree of trauma. One member of the pair was treated with an active pulsing electromagnetic field generator while the other was exposed to a placebo generator. They were not matched for pre-treatment pain, swelling, or disability. After three days of treatment the treated patients showed about twice as much recovery as the controls

We (_____ et al, 1995) conducted a double blind study using the same type of pulsing electromagnetic field generator proposed for this study to treat tibial and metatarsal stress fractures. The group exposed to actual fields improved significantly faster than those exposed to placebo fields.

Kaplan & Weinstock (1968) performed a double blind study with 100 foot surgery patients and found that pulsed fields reduced pain and edema more than placebo treatment among 100 patients undergoing foot surgery. Unfortunately, their ratings were subjective so we can not be certain that the rating scales were applied uniformly. The technique has been successfully used to prevent initial development of edema and pain in burn patients (Ionescu et al 1982). It has also been successfully used to reduce swelling and control pain among 250 patients with non-operative hand injuries participating in a controlled study (Barclay et al 1983).

Uncontrolled clinical trials have reported the use of low frequency pulsing electromagnetic fields to speed and promote the healing of delayed union and nonunion fractures since the 1970s (e.g. Sharrard 1989). At least 14 of the papers report the technique's use for these problems in the tibia. Taken together, they represent trials with 1,275 patients of whom an average of 81% healed after a significant pause in progress (Technology Evaluation, 1989). More recently, double blind studies indicating the technique's effectiveness on a wide variety of bones have been published. For example, Sharrard (1989) performed a double blind study of 45 fractures of the tibial shaft and in which 20 received active coils and 25 received dummy units. Orthopedic examination indicated that nine of the subjects in the active group showed healing relative to three in the

control group. Objective radiological evaluation indicated union of five fractures and progress toward union in an other five in the active group compared with union in one fracture and progress toward union in one fracture in the control group. Salzberg et al (1995) recently performed a randomized, double-blind study in which they treated pressure ulcers on spinal cord injured patients with either real or placebo pulsing electromagnetic fields. The group of 20 patients receiving real pulsing electromagnetic fields showed 84% of grade II ulcers healed within one week relative to 40% for the placebo group. The real pulsing electromagnetic field group required an average of 13.0 days to heal vs. 31.5 days for the placebo group. Duma-Drzewinska and Buczynski (1978) did an uncontrolled trial with 27 patients having bed-sores and found that those with superficial ulcers healed more quickly than expected. Itoh et al (1991) performed an uncontrolled clinical study of twenty-two patients with stage II diabetic ulcers unhealed for between three and twelve weeks or stage III ulcers unhealed for between eight and 168 weeks. When pulsing electromagnetic field therapy was added to ongoing traditional therapeutic approaches, all twenty-two ulcers healed. The stage II ulcers healed in between one and six weeks and the stage III ulcers healed in between one and twenty-two weeks.

There are no published studies on the use of pulsing electromagnetic field for treatment of headache. However, the "Magneto-therapy" arena of alternative medicine consistently includes treatment of migraine headaches as one of its uses (e.g. Newman 1995, Washnis and Hricak, 1993).

(3) Relationships between extracranial blood flow and headache activity: Temperature biofeedback from the hands and feet has been shown to increase peripheral blood flow (Freedman 1991, Iezzi, Adams, and Sheck, 1993). Double-blind and clinical studies have shown that this training results decreased migraine headache activity which is sustained for ten to fifteen years (Blanchard and Andrasik 1985, Iezzi, Adams, and Sheck, 1993). The chain of events initiated by increasing peripheral blood flow which result in this sustained decrease in headaches is not understood (Reich and Gottesman, 1993) but it is known that changes in extracranial blood flow do influence headache activity for about half of the migraine headache patients studied (Gillies and Lance, 1993). It is possible that increases in peripheral blood flow caused by exposing the limbs to pulsing electromagnetic fields produce any effects they may have on migraine headache activity by initiating a chain of events similar to that initiated by temperature biofeedback training. If migraine headaches are mainly related to vascular problems and tension headaches are mainly related to muscle contraction problems, only the migraine headache patients should respond to exposure to pulsing electromagnetic fields. It is reasonable for headache activity to return to pre-treatment baseline levels within a few months after cessation of treatment as there is no reason for changed blood flow patterns to be maintained. Thus, the treatment would have permanent effects only with monthly (or so) refreshers provided by a home unit. If this portion of the hypothesis is confirmed, the rate of return to baseline headache activity will help determine the design of a follow-up study intended to determine the optimal inter-treatment interval. Of course, pulsing electromagnetic fields may produce their effects through some unknown effect not related to blood flow at all.

(4) Evaluation of headache pain and intervention effectiveness:

(a) Visual - analog pain rating scale: The scale will be used with our patients to help them rate the intensity of their pain. It consists of a colored bar with numbers under it. The bar is white on the left end and gradually changes through darker shades of pink to deep red at the right end. The numbers go from zero under the white through ten under the deep red. The words "no pain" are printed to the left of the zero and white area of the bar and the words "maximum pain" are printed to the right of the "10" and deep red area of the bar. The patient is shown the scale and told that zero indicates no pain and that ten indicates the most pain imaginable. The patient looks at the scale and gives a number representative of the pain intensity at that moment or over the time period requested by the interviewer. The visual-analog scale has been in use for many years and has been determined to be the most reliable and valid way to assess changes in an individual patient's pain intensity across numerous evaluation sessions. Reliability and validity have been reviewed by Huskisson (1983). He notes that correlations between successive measurements of pain have been as high as 0.99 and are usually at the level.

(b) Requirement for a placebo control group: Pain is very reactive to placebo intervention (e.g. Beecher's pioneering study in 1955) so realistic placebos are a requisite part of evaluating any new intervention (Cooper 1981). A placebo / non-specific effects control group is vital to the study design because headache studies usually find about a thirty percent, short term response to inactive interventions. For example, Couch (1993) reviewed twelve placebo controlled headache studies and found a range of placebo response from four to fifty-five percent with most in the thirty percent range. While most studies, including those reviewed by Couch, use medicinal placebos, machines have been shown to produce effective placebo responses as well (Schwitzgebel and Traugott, 1968). Our study would be especially likely to produce a placebo response because of the impressive nature of the device itself and the intense □treatment□ regime which requires patients to make twenty visits to a major medical center. Non-specific effects would also be highly likely as all participants take time out of their normal routines to sit quietly in a comfortable room away from their daily stresses for an hour per day.

Evaluation of headache activity: Treatment success is usually defined as at least a 50% decrease in headache activity based on frequency, duration, and intensity with a commensurate decrease in medication use (Blanchard and Andrasik 1985). Subjects in headache studies usually keep a daily log of the frequency, duration, and intensity of headaches as well as use of headache related medications before, during and after the intervention period(s). Subjects usually rate their pain on a visual analog pain scale such as the one discussed above. The efficacy of logs (sometimes called diaries or daily charts) for tracking headache activity is very high (McKee 1993, Blanchard and Andrasik 1985). We have been using these logs for both research and clinical patients since our early □alternative medicine□ studies on biofeedback (e.g. Sherman, 1982).

b. Significance:

(1) Medical Significance: Most adults in the United States have at least occasional headaches. Headache is now the leading medical cause of lost days of work and costs the U.S. many billions per year to treat. The Nuprin Pain Report (1987) found that 157 million work days per year were lost due to this problem alone. A large minority of patients with migraine headaches are not adequately controlled with current treatments. Many of the effective treatments have significant side effects and require life-long drug therapy. This study is a direct effort to determine whether a technique already in use for other problems can effect migraine headaches as well. If it can, a new technique would be added to the currently flawed armamentarium. Its demonstrated lack of side effects makes it especially attractive.

(2) Scientific significance: Many theories concerning the underlying mechanisms producing migraine headaches have collapsed under the weight of negative evidence. However, it is known that effective treatments such as temperature biofeedback for warming the fingers do cause an increase in peripheral vasodilation. Pulsing electromagnetic fields cause an increase in peripheral vasodilation through peripheral mechanisms. If it has a similar effect on migraine headaches as temperature biofeedback, then we will have some hints about physiological changes which alter migraines.

If the technique effects muscle tension headaches the same way it effects migraines, the two descriptive types of headaches are not likely to have entirely different underlying mechanisms, although they might still be triggered by different processes. If the technique has differing effects, the underlying mechanisms may differ. This would have important connotations for the theoretical classification of headaches.

4. PRELIMINARY STUDY

An open trial exposing subjects with migraine headaches to pulsing electromagnetic fields: Eleven subjects with long histories of having poorly controlled migraine headaches at least once per week kept a one month headache log before and after exposure to pulsing electromagnetic fields. Each was exposed to pulsing electromagnetic fields for between two and three weeks for one hour per day, five days per week. The generating device described above (Diapulse model D103) was used at a setting of 975 watts pulsing 600 times per second. The field was applied above each femoral artery at the medial quadriceps (inner thigh). This site was chosen because it produced the greatest increases in peripheral blood flow of all the sites we tried. Headache activity was tracked daily by the therapist during the exposure phase. All subjects discontinued use of prophylactic medications after the initial two week baseline before being exposed to the fields. Additional follow-ups beyond the two week post-treatment period were available for three of the subjects with two being followed for an additional four weeks and one for an additional five weeks. A fourth subject was followed for an additional two weeks but became ill just after the end of treatment.

The results for each individual are presented in Table 1 and the statistical differences between pre and during treatment periods are presented in Table 2. The average number of headaches per

week decreased from 4.2 during the baseline period to 1.1 during the exposure period

Table 1 - (Already presented in Sample 5 so not repeated here.)

Table 2

Statistical evaluation of the data from the migraine headache pilot

p = probability, SD = standard deviation, t = paired t test, W = Wilcoxon test, F= repeated measures analysis of variance coefficient.

Table 2		BEFORE TREATMENT	DURING TREATMENT	AFTER TREATMENT
NUMBER OF MIGRAINE HEADACHES PER WEEK	X	4.2	1.1	0.6
	SD	2.2	2.0	2.1
	t	Pre to during = 4.8	During to post = 3.6	Pre to post =
	DF	10		
	P	0.001	0.005	0.001
	F / P	p= 0.001, F = 24.7		
NUMBER OF MILD HEADACHES PER WEEK	W	13		
	P	0.06		
NUMBER OF MODERATE HEADACHES PER WEEK	W	36		
	P	0.06		
NUMBER OF SEVERE HEADACHES PER WEEK	W	28		
	P	0.06		

(statistically different at $p=0.001$; paired “ t ”= 5.7 with 7 DF). The average number of headaches continued to decrease 0.6 during the two week post exposure period. This rate is significantly different from the exposure period ($p= 0.005$; paired “ t ”= 3.6 with 10 DF) as well as the pre-exposure baseline period ($p= 0.001$; paired “ t ”= 5.12 with 10 DF). A one way, repeated measures analysis of variance indicates that there was an overall difference between the periods ($p = 0.001$, $F= 24.7$). Patient six had a sinus infection while overseas during the two week follow-up period. She also had an allergic reaction to one of the medications she took for the infection. She reported having several headaches related to the sinus infection throughout the four week period which were very different in nature than her usual migraines. Patient ten's headaches began three years before entering the study subsequent to a mild closed head injury. After completion of treatment, sinus CT showed her to have maxillary sinusitis. She was the only subject who had a traumatic onset of headaches and was also the only subject who did not have a substantial decrease in the frequency of headaches during exposure to PEMFs. Thus, it is possible that her headaches have a different underlying pathology than the others' and that the pathology is not responsive to increased peripheral blood flow.

All three patients for whom longer follow-up information was available informally reported the return of headache activity.

1. Patient three was followed for a total of six weeks after the end of treatment. At the end of the sixth week, she had a headache associated with a sharp change in barometric pressure.

2. Patient four was initially followed for six weeks during which she remained headache free. She subsequently contacted us six months after the end of treatment and reported headache frequency to be about once every three weeks with the headaches being shorter and milder than pre-treatment. Her pre-treatment baseline was a 21 year history of headaches occurring about twice per week.

3. Patient five was followed for seven weeks after the end of treatment with headaches returning to their baseline level of frequency and intensity.

Patients with long histories of migraine headaches who were exposed to pulsing electromagnetic fields showed an almost complete cessation of headache activity during the weeks of exposure and for several weeks thereafter. It is possible that exposure to pulsing electromagnetic fields did cause sufficient increase in peripheral blood flow to temporarily effect headache activity. This increase would have initiated some chain of psychophysiological events which actually caused the improvement. The other possibilities for how the effect could have been produced include a powerful placebo effect and the induction of currently unrecognized physiological effects of the fields, having nothing to do with blood flow, which somehow decreased headache activity.

In spite of the impressive nature of the PEMF generator, the investigators do not feel that the major effects were due to placebo responses because (a) the participants each had multi-year histories of unsuccessful treatments with numerous highly touted therapeutic approaches, (b) the change in headache activity was much greater than would be anticipated due to a placebo response, (c) the rate of headache activity among the few subjects followed beyond the one month follow-up increased within a month of ending exposure rather than lingering as would be anticipated of a placebo response, and (d) the rebound effect from stopping the prophylactic

medicines should have hit sometime near the end of □treatment□ so would have overwhelmed any placebo effect. Additional, but highly indirect, support for the projected mechanism of action and against the basic response being due to a placebo effect is provided by the one patient who showed minimal response to the intervention. This was the one participant who had a traumatic origin for her headaches and was found to have sinusitis so would not have been expected to respond particularly well to interventions involving increased peripheral blood flow.

Because of the powerful effects demonstrated in this trial, the investigators feel that it is worth performing a double-blind, controlled studies to determine whether this intervention is actually effective. This work was presented at the _____ annual meeting and has been submitted for publication in the International Journal of _____.

5. RESEARCH DESIGN AND METHODS

a. Design: The study design is illustrated in Figure two. This is a typical double blind, placebo controlled study with initial baseline and brief follow-up periods. Patients will be diagnosed by the participating neurologist as having migraine headaches with aura and then be randomized into real or placebo exposure groups. A one month initial baseline, during which subjects will keep a daily log of their headache activity, will be followed by half the subjects receiving actual exposure to pulsing electromagnetic fields and half receiving placebo exposure for two weeks. This is followed by a one month follow-up during which subjects keep the log again.

Figure 2

Study Structure

1. Patients are diagnosed by the participating neurologist as having migraines with auras.
2. Patients are randomized to real or placebo pulsing electromagnetic field therapy by picking an envelope from a basket.
3. Patients keep a one month log of headache intensity, duration, and frequency as well as of medication use.
4. Two weeks of daily pulsing electromagnetic field therapy (real or placebo)
One set of patients uses device □A□ and one uses device □B□. Only the PI knows which is

the placebo.

Power set at 975 peak watts and 600 CPS.

Five, one hour sessions per week with pulsing electromagnetic field aimed at the inside of the thigh (over the medial quadriceps targeting the femoral artery).

Patients provide all information they would have put into a log every time they see the therapist during a pulsing electromagnetic field session.

5. Patients keep a one month log of headache intensity, duration, and frequency as well as of medication use.

b. Subjects:

(1) Inclusion and exclusion criteria: Patients of either sex eligible for care at the Sample Institute who are between the ages of 18 and 70 and have at least a four year history of migraines with auras at least once per week. They must not have medical problems which would influence changes in blood flow (to avoid preventing the intervention from changing blood flow) or psychological diagnoses (to avoid influencing reports of pain intensity) will be invited to participate in the study. The age limits are due to agency restrictions on the lower age limit on people who can participate in placebo controlled studies at any time and on the upper age limit of people who can be treated at this medical center during the next year due to budget restrictions. There have been no reports of the generator having any effect on pregnancy, but in order to avoid potential problems, women who are in the childbearing age range who wish to participate will be required to have a urine test showing that they are not pregnant prior to beginning participation and must agree to use a standard, accepted method of birth control during the study. If a participant becomes pregnant during the study she will have to drop out immediately. No patients with medicine rebound, sinus, cluster, or other types of headaches will be able to participate. Headache diagnoses will all be made by the participating neurologist during an initial interview according to the standard International Ad Hoc Committee classification (Olesen 1994).

(2) Number of subjects: A power analysis (Cohen 1988) of the pilot data indicates that twenty subjects will be required per group ($\alpha = 0.05$, $\beta = 0.2$) to differentiate between the actual vs. placebo exposure conditions assuming a thirty percent placebo response and a ninety percent response among classical migraine subjects receiving actual exposure. The data presented in the introduction demonstrated that sufficient subjects are available.

(3) Source of subjects: Subjects will be drawn from (a) the pool of patients referred to the Neurology Clinic at Sample Institute, (b) posters placed in Family Practice, the pharmacy, and in Ob-Gyn and, (c) notices placed in the post newspaper.

(4) Subject identification: Each subject's data will be given a sequential group code when stored outside of the medical record. Clinical records will be kept in the usual way. Additional

information recorded for study purposes will be kept in a locked file until patient identification is removed and coding is substituted.

(5) Precautions and corrective actions: If a patient has an unanticipated, negative reaction to the stimulator, the person will stop using it but the results will be included in the analysis. If a participant becomes pregnant, she will have to drop out of the study immediately.

(6) Subject Privacy: During the study, all data identifiable with an individual subject will be kept in a locked file cabinet in Research Service. After each subject has completed participation, their name will be removed from all data kept in paper files and a code number will be substituted. The code will only be available to the principal investigator for use if follow-up studies are approved. All data entered into computers will have only the code numbers. At the end of the study the data will be kept only as computer records without names or other identifying information as required. Patients will be told that data from their participation is likely to be published and presented in professional forums but that their individual identity will not be released under these circumstances. They will also be told that numerous official agencies have the right to inspect research records for purposes of verifying compliance with regulations but that these agencies will not make the participants' identities public.

c. Procedure:

(1) Randomization: When subjects meet the entrance criteria, they will be diagnosed by the neurologist then randomized into actual or placebo pulsing electromagnetic field therapy. Randomization will be performed from a sequence of 40 □A□s and □B□s being generated by a computer algorithm. As subjects qualify for participation, they will be assigned the next letter on the list. The letter indicates which pulsing electromagnetic field generator they will use. This letter will be placed on all of the patient's research and clinical records so the code can be broken by the PI in case of an emergency. The number of codes will be the same as the number of subjects in the study so the groups will eventually be the same size. If subjects drop out, replacement subjects will receive the dropouts' letters in sequential order.

(2) Evaluation of headache activity: Subjects will keep a daily log of the frequency, duration, and intensity of headaches as well as use of headache related medications for one month before initial exposure to either the real or placebo device. Subjects will rate their pain on the visual analog pain scale discussed in the introduction. The scale goes from zero (no pain) to ten (so much pain that they would faint if they had to sustain it for one more second). The only evaluation will be the headache log kept before and after intervention. The log we use is typical of those shown to be highly efficacious (as discussed in the introduction) and requires only one entry per headache. The average intensity (on the zero through ten scale), the duration, and medications taken are recorded.

(3) Exposure to pulsing electromagnetic fields or placebo: After keeping the initial baseline week log, patients will be exposed to pulsing electromagnetic field (real or placebo) on the thigh at a power/frequency setting of 6/600 for one hour per day, five days per week for two

weeks. We direct pulsing electromagnetic fields to the thigh because we have found that we get more increase in peripheral blood flow when that site is used than from any others we have tried (see introduction). Neither the therapist nor the patient will know which group they are in. This will be accomplished by dedicating two pulsing electromagnetic field machines to the study. One will have the field generator disconnected from the circuit and the field indicator lights on the heads of both machines will be covered with opaque caps. The machines will look and sound the same when functioning. One will be marked "A" and one "B". Only the PI will know which is which and only he will perform the standard daily calibration procedure (to be sure the active machine is putting out the correct field strength). Patients can not feel the machine working so they will not be able to tell which group they are in. However, as a check, each will be asked whether they thought they were in the real or placebo group after each stage of the study. This will be done by having each rate how certain they are they received the real treatment on a scale of 0 - 10 where zero is not at all certain and ten is sure they received the real treatment. We are successfully using this type of placebo with ongoing tibial stress fracture and sprained ankle studies in which participants are randomized and exposed to either real or placebo pulsing electromagnetic fields.

The pilot showed that two weeks are sufficient time to produce any effect likely to occur. At the end exposure, patients will keep a one month follow-up log. Blanchard et al (1987) did an analysis of the stability of headache activity over time for different headache disorders in order to determine the appropriate duration of baselines for each disorder. They have shown that the duration of logs we propose are sufficient to ascertain the basal level of headache activity among patients with migraines with auras.

d. Data analysis plan:

(1) Sample Institute's Research Service will supply a PhD level clinically oriented biostatistician to supervise the data analysis (letter of agreement enclosed). No data identifiable with an individual patient will be kept outside of locked cabinets during the study. At the end of the study the data will be kept only as computer records without names or other identifying information for a minimum of 15 years (as per HHS regulations).

(2) Changes in headache activity during the exposure portion of the study: Differences in headache activity (defined above) will be determined from the log kept for three weeks before intervention, during the three weeks between intervention arms, and for three weeks after intervention as well as the daily reports to the therapist during the two, two week exposure phases. The first analysis will consider this to be a repeated measures design with two groups (the placebo vs. the real exposure). The repeated measure will be the various periods during which logs were kept and reports of headache activity presented. A probability of 0.05 will be considered significant because a one chance in twenty of the results being random far exceeds the predicted difference in clinically important rates of headache occurrence. The number of headaches per week and headache duration are parametric measures so a parametric test can be used as long as the distributions are normal and fields will produce a decrease in headache activity relative to exposure to a placebo device. Thus we are predicting not only a difference but the direction the difference will be in. We are also interested in establishing the risk of subjects

in each headache type group having headaches after exposure to real vs. placebo fields. This will give us the same information as the number of subjects who reach the success criterion defined by Blanchard and Andrasik (1985) of a 50% decrease in headache activity. The risk difference for each group will be calculated in accordance with Overvad's (1994) formula.

6. FUNCTIONS OF PERSONNEL:

(a) James E. Expert, PhD (Psychophysiology) is Director of Sample Institute's headache clinic. He will be responsible for directing the overall study, coordinating the data reduction efforts, and for supervising the technician. He will also coordinate with the statistician provided by Sample Institute's Research Department to insure the correct analysis of the data. His entire salary is paid by the institution so no funds are requested from the granting agency. He is both a scientist and a health care provider experienced in performing large outcome and patient evaluation studies. He has published on evaluation of headache activity and the use of pulsing electromagnetic fields and other alternative medicine techniques for headaches and other disorders. He has had grants from the Department of Veterans Affairs, NIH, the Army, and private industry to investigate headache and other pain related problems.

(b) Lucey Nervesplit, MD (Neurology) is the Neurology Associate's specialist on headache diagnostics and treatment. She will be the project's diagnostician so will see examine potential subject to insure they meet the entrance criteria. Her entire salary is paid by her institution. In April, 1996, she completed two grants from the NIH's Advanced Technology section to develop an automated headache diagnostic and database system.

© James R. Frank, DO (Neurology) is Head of Neurology Associates and has over twenty years of experience treating patients with headaches. His entire salary is paid by his institution. He will be the project's patient monitor so will follow every subject through their active participation in the study to insure that no untoward effects are unnoticed and, should any occur, are corrected immediately.

(d) Linda A. Example, BA and John Marvenite, BA are the pulsing electromagnetic field therapists for Sample Institute's Research Service. Their contracts are paid entirely from grant support and they will be the contract technicians on this project. They have several years of experience performing research studies involving pulsing electromagnetic fields including the pilot study for this project on headaches as well as similar projects on sprained ankles, diabetic ulcers, and reflex sympathetic dystrophy. Their contracts cost the service \$20 per hour. The cost includes all fringe benefits, Social Security payments, etc. This is the cost the service engenders for its other contractors performing at essentially the same level of independence and skill and is typical of costs in this area. They will (1) recruit all subjects for the study (about 80 hours), (2) track their progress through the study to ensure that they participate in all sub-sections of the study as required and as they are randomized (about 40 hours), (3) perform all 400 hour long exposures required by the study (100 hours as about 15 minutes of technician time are required per hour of exposure), (4) collect (and convince people to actually keep and then turn in) all 80 month long headache logs (about 20 hours), (5) enter all of the data into the computer (about 40 hours assuming twenty minutes per log), and (6) prepare the data for analysis as required by the

statistician provided by the research department (about 80 hours). An additional 60 hours are anticipated to account for dropouts. The study design justifies the need for two technicians and requires that approximately 500 hours (as indicated above) be dedicated to the project which will cost the service \$10,000 at the above rate of pay.

7. RESEARCH FACILITIES / ENVIRONMENT

Nine of the Institute's twenty-seven full time faculty and thirteen of the Institute's sixty-two part time faculty are actively involved in research with most conducting their studies at the Institute. Sample institute has a small but well equipped research facility which is entirely funded by grant support. Nine thousand square feet of space is dedicated to research with about half of that being laboratory space and half being testing rooms and offices for support personnel. The research facility has a full time PhD level director, a half time PhD level biostatistician, and three full time, bachelor's level technicians. Core equipment includes two fully automated neuropsychology testing batteries; three Pentium two computers with zip drives dedicated to data reduction and analysis as well as report generation; a Coulbourn physiological recording system; and seven computer based, multichannel biofeedback systems.

8. INVESTIGATORS' CONCURRENT SUPPORT

a. Dr. Expert currently has two grant supported projects in progress:

(1) Use of neurofeedback therapy to cure the common cold two weeks after onset. Supported by the non-profit foundation of the EEG equipment manufacturers' combine of North-Central Megalopolus with a \$2,751, three year grant. This project and the proposed work are unrelated. There will be no overlap of equipment or personnel other than the principal investigator.

(2) Comparative efficacy of sEMG biofeedback, progressive muscle relaxation training, Fiorinal, and placebo in preventing headaches among subjects with chronic tension headaches. Supported by NIH with a \$179,500, six year grant. This project and the proposed work are unrelated. There will be no overlap of equipment or personnel other than the principal investigator.

b. Neither Drs. Nervesplit nor Frank have active grant support at this time.

9. LITERATURE SITED

The references are essentially the same as those used for the sample protocol with a few additions for greater detail.

10 additional references with the PI as first author and one as second author were included in this list and referred to throughout the grant's text to give the grant's reviewers the idea that the investigators probably know something about the topic and have performed successful work in the field which was acceptable to recognized/mainstream, peer reviewed journals.

10. BUDGET

- a. Equipment: \$50,000
Both \$25,000 generators have been paid for by Sample Institute but will be costed against this project.
- b. Supplies: \$50
photocopying (headache logs, etc.), computer supplies, general supplies (pencils, clipboards, etc.) paid for by Sample Institute's research fund.
- c. Contract support for the two therapists required to perform the study: \$10,000 (\$5,000 each) for 250 hours each at \$20 per hour. Exact duties and hours per task were defined in the resources section.
- d. Travel to professional meeting to present results: \$400 paid for by Sample Institute.
- e. Statistical consultation: \$50 paid for by Sample Institute.
- f. Administrative processing fees and support (20% overhead fee): \$2,100 paid from Sample Institute's research fund.

TOTAL: \$12,600 (without equipment)

TOTAL WITH EQUIPMENT \$62,500

Total requested from the granting agency: \$10,000 for c above.

Section I:

An algorithm of the steps in evaluating clinical protocols and articles

Slightly adapted from an outline prepared by
Lori A. Loan, RN, PhD; Chief of Nursing Research
at Madigan Army Medical Center in Tacoma Washington
(used with her knowledge and enthusiastic consent)

1. First read the article literally to understand what the author has said. If you skim it, you are very likely to miss important information.
2. Next, read the article in a critical fashion to analyze the adequacy of the study.
 - a. Problem Statement:
 - (1) What is the problem that was studied? Is it explicitly identified?
 - (2) Is the problem stated precisely and clearly?
 - (3) Is the problem delimited in scope?
 - (4) Is the problem justified in light of theoretical and empirical work relevant to the topic?
 - (5) What does the literature say? How does the study fit with what is known? How does it contribute to gaps in knowledge?
 - (6) Is the theoretical and practical significance of the problem discussed?
 - (7) Of what importance is the problem to medical science and practice?
 - b. Conceptual Framework:
 - (1) What are the major concepts guiding the study and how are they defined?
 - (2) Are the concepts linked to one another? How?
 - (3) What theoretical perspective has been used to better understand the problem? Is a theoretical or conceptual perspective clearly identified?
 - (4) Does the conceptual framework accurately reflect the state of medical science?

c. Purpose:

- (1) What is the purpose of the study? What was the investigator trying to find
- (2) What concepts or variables are specified in the purpose? What are the
- (3) Is the purpose logically linked to the conceptual framework?
- (4) Is the purpose linked to earlier empirical work?
- (5) Does the purpose precisely indicate how the study will contribute to new knowledge? For example, will the study/article contribute description of a phenomenon, explanation of a relationship between two or more concepts, or predict an outcome?

d. Design:

- (1) What is the study design?
- (2) Is the design consistent with the stated purpose of the study?
- (3) Is the design appropriate given the state of knowledge about the topic or the current understanding of research design?
- (4) Is the design too complex or inadequate for the purpose?
- (5) Are threats to the validity of the study identified? Are they corrected where possible?
- (6) Are potentially confounding (extraneous) variables controlled by the basic design?

e. Sample:

- (1) What is the sample? Is it described clearly?
- (2) How was the sample selected? Was the method of selection appropriate given the purpose of the study? Was some type of sample bias/selection, or loss perhaps influencing the findings?
- (3) Is the sample representative of the groups to which the study findings should be applied? If not, how does it differ? What are the consequences of the difference?

f. Instruments (devices, surveys, etc):

- (1) What instruments were used to measure the outcomes / concepts?
- (2) Were they adequate reflections of the outcomes being studied? If they do a
- (3) Did the investigators show that the instruments were appropriate to measure
- (4) Do the instruments measure accurately enough for the required precision?
- (5) Are the validity and reliability of the instruments adequate for their application?
What problems would be expected with the selected instruments validity and reliability? How were they addressed?
- (6) Were the instruments appropriate for the population being studied?
- (7) If the instruments were developed for the study, what were the procedures used to assess their adequacy?

g. Procedure:

- (1) What specifically was the treatment?
- (2) Did the investigators present convincing evidence that the treatment should

- (3) What procedure was used for data production? Is it clearly described? Were the procedures appropriate? Could the methods have influenced the findings? How?
- (4) Could another investigator repeat the same study, given the description of the procedures?

h. Analysis:

(1) What procedures were used to analyze the data?

(a) First look for descriptive statistics to get a feel for what the actual original data were:

- (I) How were the variables measured (counting, scales, etc.)?
- (ii) Are there any clues as to what the distribution(s) of the major variables were like (frequencies, standard deviations, etc.)?
- (iii) Are there any implications of these descriptive data given the purpose of the study?
- (iv) Is enough raw data presented for you to get a feeling for what

were't consistently
on their own.

clinically important enough to stand

(b) Then move to the inferential statistical findings (where they are testing relationships, difference, effect or prediction):

- (I) What type of statistic is used?
- (ii) Is this the test anyone would have chosen for this design and wanted.
- (ii) Why did they use it? How does it fit in with the purpose of the study? . . . with the description, relationship, difference, effect, or prediction they wished to make?
- (iii) Every test has some element of chance connected with it. What is the probability that the results of the statistical test would be different with repeated sampling? -- usually expressed as α
- (iv) What is the meaning or clinical significance of the size of the relationship or the extend of the difference?
- (v) What are the possible explanations for the findings? Potentially confounding variables? Other internal validity issues that could have produced the results?

(2) Are the analyses described appropriate for the purpose of the study?

(3) Are the analyses appropriate for the type of data?

(4) Are negative and positive findings presented where appropriate?

(5) Are factors that might have influenced the results taken into account in the analyses?

i. Discussion:

(a) How are the findings related to previously cited research?

(b) How are the findings related to the conceptual framework?

Are the generalizations appropriate or grandiose?

(d) Are the conclusions valid and justified given how the study was done? Are they justified by the results presented?

- (e) What are the limitations of the study?
- (f) What recommendations for further study are appropriate?
- (g) What recommendations for implementing the research are appropriate? Is more work needed before the findings can be appropriately applied to practice?

3. Your overall decision about the paper:

- a. Do you trust what you read?
- b. Would you change your clinical practice based on what you read?
- c. Can you think of a way to do it better?

Section J

Can you trust the interpretation of study results by people who have not adequately tested their assumptions or are not neutral?

This section contains several examples of distortions of study results due to (a) unwarranted assumptions about so aspect of the study population, (b) huge impacts of previously unrecognized co-variables and, (c) purposeful distortions by ignorant / biased / self-serving politicians and writers acting out in the public press.

Both examples are based on studies of the genetics of violence.

Two points that people (usually commentators) interpreting the results of human behavioral genetics studies frequently miss:

1. Imprisonment does NOT equal propensity to violence!

Nearly all prison inmates are very dumb and non-violent. A disproportionate number of truly stupid people are in prison compared with general intelligence of criminals not in prison.

2. Few people are violent! Nearly all of the civilian violence in any human population is caused by about 7% of the population – who are virtually all chronically violent throughout their lives.

Genetic background of extreme violent behavior

For example, Tiihonen et al found that in Finland's nearly all Caucasian population:

“In developed countries, the majority of all violent crime is committed by a small group of antisocial recidivistic offenders.”

Most studies come up with a range of 5 to 7 percent of the population.

Goodwin led a 1992 national survey of young offenders. They “found that in youth offenses involving some form of violence, 80 percent of the offenses were committed by 7 percent of the population.” There was no correlation between violence and race at all, when you took socioeconomic status out of it--in fact, black middle-class kids, we'd previously found, were less likely to abuse drugs than white middle-class kids and were more socially responsible.”

The early day of the debate - the 1970s and 80s:

XYY in prison populations and violence

Criminality in men with Klinefelter's syndrome and XYY syndrome: a cohort study

[Stochholm et al:](#) the criminal pattern in men between 15 and 70 years of age diagnosed with 47,XXY (Klinefelter's syndrome (KS)) or 47,XYY compared to the general population.

Design Register-based cohort study comparing the incidence of convictions among men with KS and with 47,XYY with age- and calendar-matched samples of the general population. Crime was classified into eight types (sexual abuse, homicide, burglary, violence, traffic, drug-related, arson and 'others').

Setting Denmark 1978–2006.

Participants All men diagnosed with KS (N=934) or 47,XYY (N=161) at risk and their age- and calendar-time-matched controls (N=88 979 and 15 356, respectively).

Results: The incidence of convictions was increased in men with KS (omitting traffic offenses) compared to controls with a HR of 1.40 (95% CI 1.23 to 1.59, $p < 0.001$), The incidence of convictions was significantly increased among men with 47,XYY compared to controls with a HR of 1.42 (95% CI 1.14 to 1.77, $p < 0.005$).

NOTE: This study does not discuss violence but it has been misinterpreted as being about it.
Genetics and violent behavior

For example, Tiihonen et al found that in Finland's nearly all Caucasian population:

“Our results, from two independent cohorts of Finnish prisoners, revealed that a monoamine oxidase A (MAOA) low-activity genotype (contributing to low dopamine turnover rate) as well as the CDH13 gene (coding for neuronal membrane adhesion protein) are associated with extremely violent behavior (at least 10 committed homicides, attempted homicides or batteries).

No substantial signal was observed for either MAOA or CDH13 among non-violent offenders, indicating that findings were specific for violent offending, and not largely attributable to substance abuse or antisocial personality disorder. These results indicate both low monoamine metabolism and neuronal membrane dysfunction as plausible factors in the etiology of extreme criminal violent behavior, and imply that at least about 5–10% of all severe violent crime in Finland is attributable to the aforementioned MAOA and CDH13 genotypes.”

Note: the violent behaviors studied were usually purposeless and spontaneous.

More on the study: NIH-led study identifies genetic variant that can lead to severe impulsivity
A multinational research team led by scientists at the National Institutes of Health has found that a genetic variant of a brain receptor molecule may contribute to violently impulsive behavior when people who carry it are under the influence of alcohol. A report of the findings, which include human genetic analyses and gene knockout studies in animals, appears in the Dec. 23 issue of Nature.

"Impulsivity, or action without foresight, is a factor in many pathological behaviors including suicide, aggression, and addiction," explains senior author David Goldman, M.D., chief of the Laboratory of Neurogenetics at the NIH's National Institute on Alcohol Abuse and Alcoholism (NIAAA). "But it is also a trait that can be of value if a quick decision must be made or in situations where risk-taking is favored."

In collaboration with researchers in Finland and France, Dr. Goldman and colleagues studied a sample of violent criminal offenders in Finland. The hallmark of the violent crimes committed by individuals in the study sample was that they were spontaneous and purposeless.

"We conducted this study in Finland because of its unique population history and medical genetics," says Dr. Goldman.

"Modern Finns are descended from a relatively small number of original settlers, which has reduced the genetic complexity of diseases in that country.

Studying the genetics of violent criminal offenders within Finland increased our chances of finding genes that influence impulsive behavior."

The researchers sequenced DNA of the impulsive subjects and compared those sequences with DNA from an equal number of non-impulsive Finnish control subjects.

They found that a single DNA change that blocks a gene known as HTR2B was predictive of highly impulsive behavior. HTR2B encodes one type of serotonin receptor in the brain. Serotonin is a neurotransmitter known to influence many behaviors, including impulsivity.

"Interestingly, we found that the genetic variant alone was insufficient to cause people to act in such ways," notes Dr. Goldman. "

Carriers of the HTR2B variant who had committed impulsive crimes were male, and all had become violent only while drunk from alcohol, which itself leads to behavioral disinhibition.“ The researchers then conducted studies in mice and found that when the equivalent HTR2B gene is knocked out or turned off, mice also become more impulsive.

How about inheritance of a propensity to violence and delinquency?

Taylor, Iacono, & McGue found evidence for a genetic etiology of early-onset delinquency. Early onset delinquency is more genetic than late onset delinquency based on studies of 11 year old twins - 36 pairs of early starters, 86 late starters, 25 controls.)

Van den Oord, Boomsma, & Verhulst studied genetic and environmental effects on the co-occurrence of problem behaviors in three-year-old twins. Based on 446 monozygotic, 912 dizygotic pairs of 3-YO twins, 37% of behavior due to genetics, 51% due to shared environment and 11% due to non-shared environment.

Alsobrook and Pauls performed a review of twin studies and reported overwhelming evidence supporting genetic component to violent behavior among young children.

Guo et al have conducted a series of studies showing that about one percent of people have any combination of three genes giving a propensity to violence.

The VNTR 2-Repeat in *MAOA* and Delinquent Behavior in Adolescence and Young Adulthood: Associations and *MAOA* Promoter Activity.

Dopamine Transporter, Gender, and Number of Sexual Partners among Young Adults.

Contributions of the *DAT1* and *DRD2* Genes to Serious and Violent Delinquency among Adolescents and Young Adults.”

The next round in the debate after XXY was discredited:

Recent (2000s) findings about shortened *MAOA* “WARRIOR” genes

“People who have a shortened *MAOA* gene do not produce a protein needed to break down old serotonin in their brains. These people are more likely to be agitated, aggressive, and impulsive. The gene can come in the form of 2, 3, 3.5, 4, or 5 allele. A 3-repeat allele is considered dysfunctional and is what is referred to as the “warrior gene”. A 2-repeat allele is considered very dysfunctional. People with a 2-repeat allele *MAOA* gene have a permanent chemical imbalance in their brain making the person more likely to be agitated, aggressive, and impulsive. According to a study published in *Comprehensive Psychiatry*, .5% of whites have the 2-repeat allele version compared to 4.7% of blacks.

That means blacks are 9.4 times more likely to have the extremely dysfunctional version of the gene than whites. Considering that black Americans are 9 times more likely to commit murder, this is very significant.

Other studies have shown even higher rates of occurrence of the 2-repeat allele version of the gene in blacks. Other studies show that the 2-repeat allele version is almost completely non-existent in Asians.”

The modern debate began in the 1990s:

Violence, Genes, and Prejudice

By [Juan Williams](#) Discover November 01, 1994

Can genes make one person more likely to act violently than another? Can the question even be asked in a country where violence--in many people's eyes--has come to wear a young black face? As scientific debates go, the war of words over the genetic roots of violence has itself been marked by unusual violence. It has damaged careers, provoked comparisons with Nazi pogroms, and prompted bitter talk of science being corrupted by political correctness. It has also sparked passionate statements about racists, Luddites, and monkey sex. This is the stuff of great fiction.

Among the more enraged critics of the entire concept is Samuel Yette, an author and former Howard University journalism professor. Yette, who is black, told the Chronicle of Higher Education that a conference on the topic would encourage the impression that blacks are born criminals.

Wasserman

“Let’s leave aside for the moment the question of whether a convincing connection can yet be made between certain genes and violent behavior. Even without conclusive evidence that it can, heated questions are being raised. Will the government try to screen people to see if they have genes that incline them to violence? If people do have such a gene, can they be forced into medical therapy? What if tests are used selectively to screen minority children, on the grounds that a growing number of American prison inmates are black or Hispanic? "Research into genetic factors has tremendous impact, and it is likely to yield controversial findings that are highly susceptible to abuse and misunderstanding," says David Wasserman, who teaches philosophy of law, medicine, and social science at the University of Maryland's Institute of Philosophy and Public Policy. A 1992 conference Wasserman planned on "genetic factors in crime" had its federal funding yanked after it was denounced for fostering racial prejudice and promoting a "modern-day version of eugenics." Research presented at the conference, its more vehement opponents protested to the New York Times, "would inevitably target minority children in the inner city in the guise of preventing future crime." So, what do we do with all this?

XYX = dumb and suggestible but not violent

Violence genes - in the same 7% that are violent????

Why more common among blacks?

So what?

The key here is that blacks are not more violent than Caucasians, Arabs, Asians, etc.

Just look at the recent “ethnic cleansing” in central Europe and the current Arab vs. all nightmare.

Section K

Use of Effect Size Calculations to determine efficacy of biofeedback based interventions

This section contains detailed evaluations of the efficacy of biofeedback based interventions for (1) prevention of tension headaches and migraine headaches of non-traumatic origin, (2) jaw area pain due to muscle problems (TMD), (3) ADHD among children, (4) Anxiety, (5) Rayndauds syndrome, (6) Urinary incontinence among adult women due to muscle tension problems, (7) Chronic Pain, (8) Epilepsy, and (9) Functional constipation.

1. Headache Pain Emphasizing Temperature and Muscle Tension Biofeedback for Prevention of Tension Headaches (including jaw area musculoskeletal pain) and Migraine Headaches of Non-Traumatic Origin

(Much of the following material is summarized from Sherman 2013 and Andrasik 2012.)

Temperature and Muscle Tension Biofeedback for Prevention of Tension Headaches (including jaw area musculoskeletal pain) and Migraine Headaches of Non-Traumatic Origin have been shown to be superior or equal to preventive medications for prevention of (non-traumatic origin) migraine and tension headaches in overall effectiveness, lack of side effects, and duration of effect.

Here is a summary of the effectiveness of muscle tension and temperature feedback for treating non-traumatic origin migraine and tension headache:

- 15 year follow-ups – if it works, patients stick with it.

 - Not true for medications.

- Controlled studies with over 3,500 participants

- Large groups

- Comparative effectiveness studies

- 7 Medical groups including American College of Neurology now recommend biofeedback to be first line of treatment for children with headaches.

The following material is from a presentation Frank Andrasik did at AAPB's 2012 Portland (Oregon) meeting which summarized the crucial data on this topic.

He has been kind enough to permit me to use modified versions of his slides for this portion

of the paper.

Evidence Base for determining efficacy:

Efficacy Reviews

1. • Qualitative
 - Panel of experts
 - Rigorous design criteria
 - Consensus reached
2. Meta-Analytic Reviews
3. Quantitative/Statistical review

Efficacy Panels

- National Institutes of Health (US)
- Diagnostic & Therapeutic Technology Assessment (JAMA)
- Canadian Headache Society
- Clinical Psychology Division of APA
- US Headache Consortium
- 7 Physician Societies
- Society of Pediatric Psychology
- Association for Applied Psychophysiology & Biofeedback
- Cochrane Collaboration

US Headache Consortium recommendations:

Relaxation training, thermal biofeedback combined with relaxation training, EMG biofeedback, and cognitive-behavioral therapy may be considered as treatment options for prevention of migraine (Grade A Evidence)

- Behavioral therapy may be combined with preventive drug therapy to achieve additional clinical improvement for migraine (Grade B Evidence)
- Evidence-based treatment recommendations are not yet possible regarding the use of hypnosis, acupuncture, TENS, cervical manipulation, occlusal adjustments, hyperbaric oxygen (Grade C Evidence)

Interpreting the results of meta-analyses of migraine treatments:

Effect Sizes for treatment of Migraines

Small effect – effect size of .2 - .5

Moderate effect – effect size of .5 - .8

Large effect – effect size greater than .8.

From: Nestoriuc Y, Martin A, Rief W, Andrasik F. (2008). Biofeedback treatment for

headache disorders: A comprehensive efficacy review. *Applied Psychophysiology and Biofeedback*, 33, 125-140.

Data on efficacy of BFB for HA compiled from:

94 Studies, 3,500 Patients

- 56 Migraine, Mean 40 Patients per study
- 45 TTH, Mean 29 Patients per study
- 7 Both HA types
- Results same for Intention To Treat Analyses (LOCF)
- Results held at FUP, Mean 14 months
- “Fail Safe Analyses”/Bias Potential
- >4,000 studies with 0 effects to reduce mean effect score to 0.00
- 148 migraine studies with 0 effects to reduce mean effect score to small (0.20)
- 168 TTH studies with 0 effects to reduce mean effect score to small (0.20)

Meta-Analyses for efficacy of BFB for Migraine HA

Blanchard, Andrasik et al. (1980)

- Holroyd et al. (1984)
- Blanchard & Andrasik (1987)
- Holroyd & Penzien (1990)
- Haddock, Rowan, Andrasik et al. (1997)
- Goslin et al. (1999)
- Eccleston et al. (2002)
- Nestoriuc & Martin (2007)
- Nestoriuc et al. (2008)

Goslin et al., Tech Rev 2.2, AHCPR, 1999.

McCrary et al. 2001

Eccleston et al. *Pain* 2002.

Nestoriuc & Martin. *Pain* 2007.

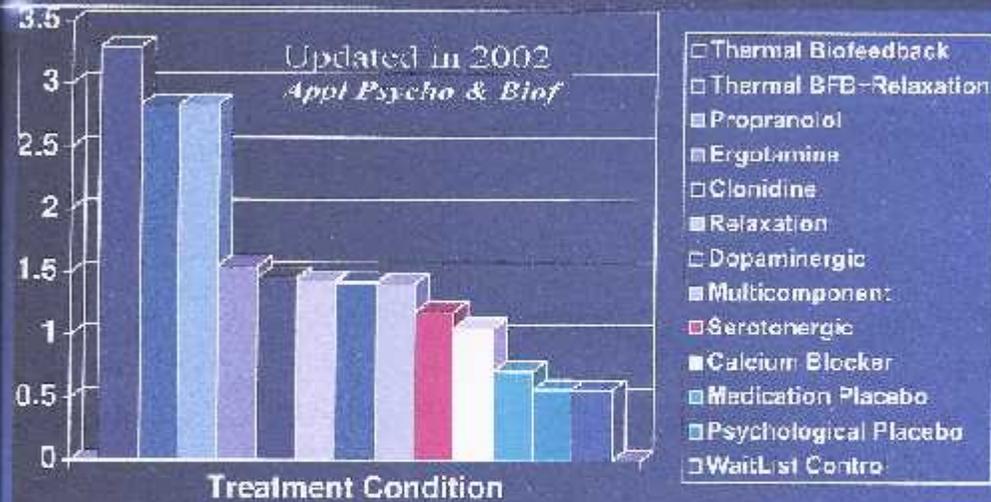
Nestoriuc et al. *Appl Psycho Biof*, 2008.

Effect sizes for tension headaches

Nestoriuc Y, Martin A, Rief W, Andrasik F. (2008). Biofeedback treatment for headache disorders:

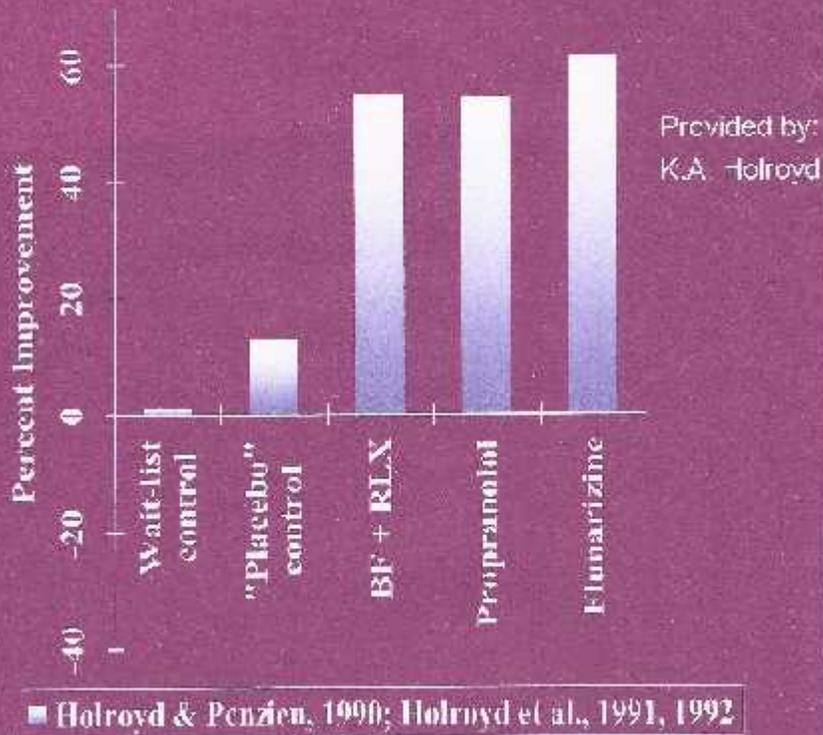
A comprehensive efficacy review. *Applied Psychophysiology and Biofeedback*, 33, 125-140.

ES Values for Nonpharmacological (n=17) & Pharmacological (n=24) Treatment (Petersmann, et al., Pain, 1996)



2003_AA-29_Portland

Meta Analysis of Behavioral vs. Drug Therapy for Migraine



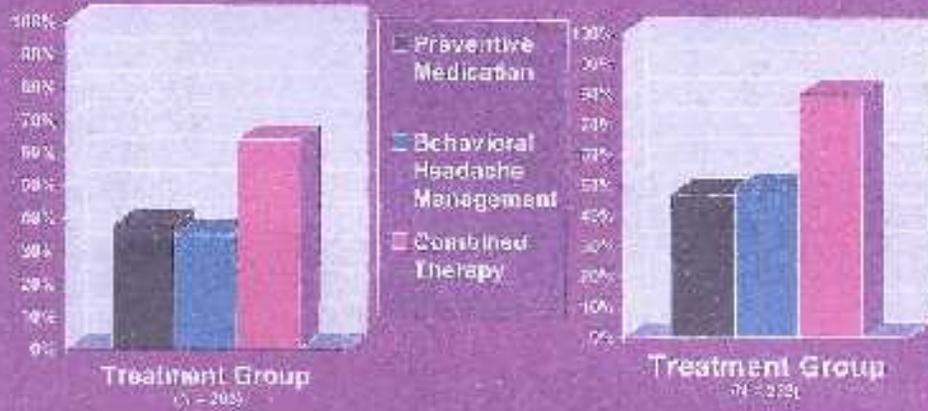
Behavior Management Enhances Drug Therapy Outcomes

(Provided by K.Z. Holroyd, Ph.D.)

% Patients \geq 50% Improved

Chronic Tension Headache
(N = 25 Arms)

Frequent Migraine
(N = 25 Arms)

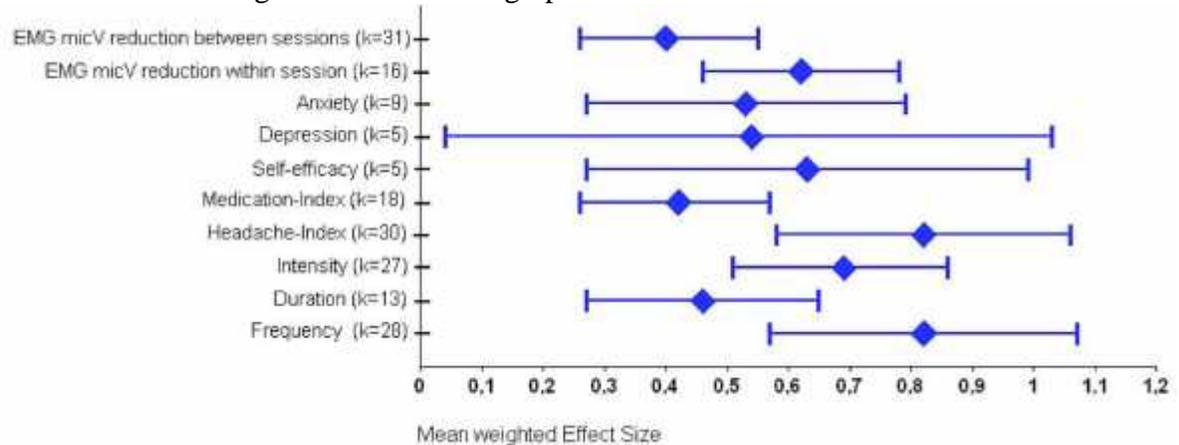


JAMA (2008)

Two NIH Trials

BMJ (2010)

The following graph summarizes the results of meta-analyses indicating changes in various symptoms related to tension headaches. The effect sizes plus and minus one standard deviation are shown along the bottom of the graph.



Meta-Analyses for efficacy of BFB for Tension HA Tension-Type Headache

- Blanchard, Andrasik et al. (1980)
- Holroyd & Penzien (1986)
- Bogaards & ter Kuile (1994)
- Haddock, Rowan, Andrasik et al. (1997)
- McCrory et al. (2001)
- Eccleston et al. (2002)
- Nestoriuc et al. (2008)

Blanchard, Andrasik et al. *Beh Ther* 1980.

Holroyd et al., paper, AASH, 1984.

Holroyd, Penzien. *J Behav Med* 1986.

Blanchard, Andrasik. In *Biofeedback: Studies in clinical efficacy*. 1987.

Holroyd, Penzien. *Pain* 1990.

Bogaards, ter Kuile. *Clin J Pain* 1994.

Haddock, Rowan, Andrasik et al. *Cephal* 1997

(From RS: This is a crucial slide as it shows how effective behavioral treatments are for migraine HA relative to (especially preventive) medications.

(end of Andrasik's material)

What about EEG and HRV feedback for HA?

While other biofeedback interventions for headache, especially HRV and EEG BFB, show great potential, there is as yet insufficient evidence to recommend applying them as initial treatments for well diagnosed tension and non-traumatic origin headaches. You certainly should not charge for them.

The studies showing efficacy at levels required by the current ethical and legal environment

simply have not been done yet.

What about biofeedback for other types of headaches? Many biofeedback interventions show great potential for treating many types of headaches but the studies showing efficacy at levels required by the current ethical and legal environment simply have not been done yet.

There is no good evidence that behavioral interventions help cluster headaches, trigeminal headaches, TMJ (the joint problem), etc.

2. Temporomandibular Disorder (TMD)

Level 4: Efficacious

Used alone, biofeedback improves pain, pain-related disability, and mandibular functioning (Gardea, Gatchel, & Mishara, 2001). When used in combination with other treatments, such as intraoral

applications (Turk, Zaki, & Rudy, 1993), and in cognitive-behavioral skills training (Gardea et al. 2001),

the effect is enhanced (Turk, Rudy, Kubinski, Zaki, & Greco, 1996). A meta-analysis of 13 studies of

EMG biofeedback treatment showed biofeedback was superior to no treatment or psychological placebo

control for patient pain reports, clinical exam findings, and/or ratings of global improvement (Crider &

Glaros, 1999).

Gatchel, Stowell, Wildenstein, Riggs, and Ellis (2006) conducted a randomized clinical trial to evaluate the efficacy of a biopsychosocial intervention for patients who were at high risk (HR) of progressing from acute to chronic TMD-related pain. The authors assessed pain and psychosocial measures at intake and at one-year follow up. Two conditions were studied: standard care and standard

care plus CBT and biofeedback comprised of frontal EMG and finger temperature training. Of 101

subjects who started the study, 98 completed the one-year follow-up study. Subjects' self-reported pain

levels were measured on an analog scale and as a response to palpation. At one year, the treatment group

subjects had significantly lower levels of self-reported pain and depression. The normal treatment group

subjects had utilized more health care for jaw-related pain. The normal treatment group subjects were

12.5 times as likely to have a somatoform disorder, more than seven times as likely to have an anxiety

disorder, and 2.7 times more likely to have an affective disorder at one year compared with treatment

group subjects.

In a recent review of the literature, Crider, Glaros, and Gevirtz (2005) report on 14 controlled and

uncontrolled outcome evaluations of biofeedback-based treatments for TMD published since 1978. This

literature includes RCTs of three types of biofeedback treatment: 1) surface electromyographic (SEMG) training of the masticatory muscles, 2) SEMG training combined with adjunctive cognitive-behavioral therapy (CBT) techniques, and 3) biofeedback-assisted relaxation training (BART). Based on a detailed review of RCTs supplemented with information from nonRCT findings, the authors concluded SEMG training with adjunctive CBT is an efficacious treatment for TMD, and both SEMG training as the sole intervention and BART are probably efficacious treatments. Medicott and Harris (2006) reported the results of a systematic review of the effectiveness of exercise, manual therapy, electrotherapy, relaxation training, and biofeedback in the management of TMD. Thirty studies met four criteria: 1) subjects were from one of three groups identified in the first axis of the Research Diagnostic Criteria for TMD, 2) the intervention was within the realm of physical therapy practice, 3) an experimental design was used, and 4) outcome measures assessed one or more primary presenting symptoms were found. Among other recommendations, the authors state combinations of active exercises, manual therapy, postural correction, and relaxation techniques often combined with biofeedback may be effective. In another recent systematic review, Turp et al. (2007) found 11 RCTs that met the criteria of at least four weeks of interventions where simple therapy was compared to multimodal interventions. Their conclusions were that with patients with no psychological disturbances simple treatment is effective, but for those with comorbid conditions a multimodal program is needed. Myers (2007) reported on a systematic review to TMD treatments and, based on a collection of previously reviewed studies and yet-to-be-reviewed studies, concludes biofeedback has been shown to be consistently superior to placebo or no-treatment controls. However, when compared to other treatments, biofeedback had mixed results: sometimes superior, sometimes equivalent, and sometimes less effective.

References

- Crider, A.B., & Glaros, A.G. (1999). A meta-analysis of EMG biofeedback treatment of temporomandibular disorders. *Journal of Orofacial Pain*, 13(1), 29-37.
- Crider, A., Glaros, A.G., & Gevirtz, R.N. (2005). Efficacy of biofeedback-based treatments for temporomandibular disorders. *Applied Psychophysiology and Biofeedback*, 30(4), 333-345.
- Gardea, M.A., Gatchel, R.J., Mishra, K.D. (2001). Long-term efficacy of biobehavioral treatment of temporomandibular disorders. *Journal of Behavioral Medicine*, 24(4), 341-59.
- Gatchel, R.J., Stowell, A.W., Wildenstein, L., Riggs, R., & Ellis, E., 3rd. (2006). Efficacy of an

early

intervention for patients with acute temporomandibular disorder-related pain: A one-year outcome study.

Journal of the American Dental Association, (1939), 137(3), 339-347.

Medlicott, M.S., & Harris, S.R. (2006). A systematic review of the effectiveness of exercise, manual

therapy, electrotherapy, relaxation training, and biofeedback in the management of temporomandibular

disorder. *Physical Therapy*, 86(7), 955-973.

Myers, C.D. (2007). Complementary and alternative medicine for persistent facial pain. *Dental Clinics*

North America, 51(1), 263-274.

Turk, D.C., Rudy, T.E., Kubinski, J.A., Zaki, H.S., & Greco, C.M. (1996). Dysfunctional patients with

temporomandibular disorders: Evaluating the efficacy of a tailored treatment protocol. *Journal of Consulting Clinical Psychology*, 64(1), 139-46.

Turk, D.C., Zaki, H.S., & Rudy, T.E. (1993). Effects of intraoral appliance and biofeedback/stress

management alone and in combination in treating pain and depression in patients with temporomandibular disorders. *Journal of Prosthetic Dentistry*, 70(2), 158-64.

Turp, J.C., Jokstad, A., Motschall, E., Schindler, H.J., Windecker-Getaz, I., Ettl, D.A. (2007). Is there a

superiority of multimodal as opposed to simple therapy in patients with temporomandibular disorders? A

qualitative systematic review of the literature. *Clinical Oral Implications Research*, 18(Suppl. 3), 138-150.

3. EEG Biofeedback for Attention Deficit Hyperactivity Disorder (ADHD) in children

The following material is largely based on (a) "Evidence-Based Practice in Biofeedback" by Yucha, C and Montgomery, D. AAPB, 2008 and (b) Arms, M et al "Efficacy of Neurofeedback Treatment in ADHD: The Effects on Inattention, Impulsivity, and Hyperactivity: A Meta-Analysis. *Clinical EEG and Neuroscience* 40: 180, 2009.

The following figure is from the Arms article referenced above. It provides an excellent summary of the effect sizes generated by both controlled and pre-post studies. The effect sizes are well within the range anticipated for good treatments.

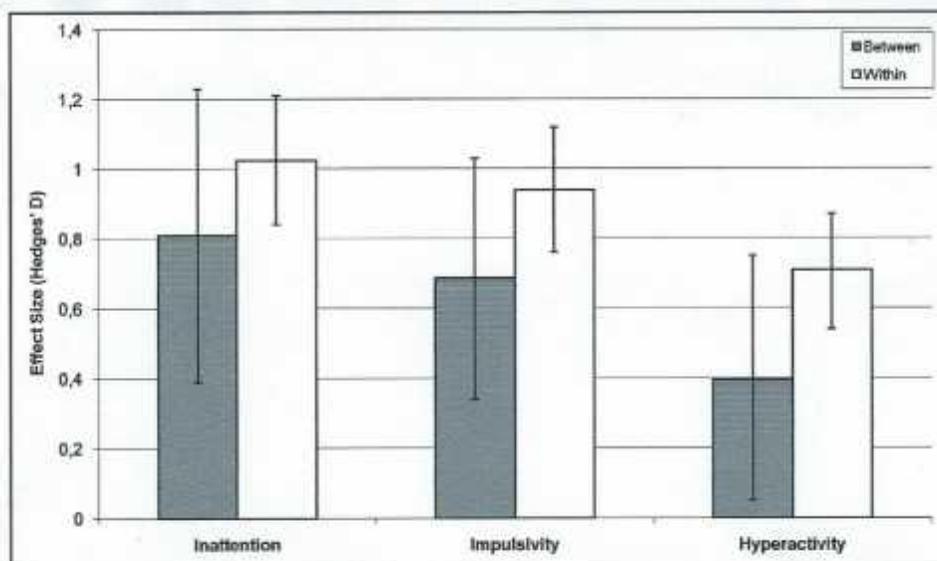


Figure 3. This figure shows the grand mean ES for the controlled studies compared to the within-subject effect sizes for all studies for all 3 core symptoms. Note that the ES for the controlled studies are slightly smaller, which could be due to the fact that many controlled studies used a "semi-active" control group. Furthermore, given the 95% confidence intervals the ES for inattention, hyperactivity and impulsivity are significant for both comparisons.

From Yucha, et al, 2008: Level 4: Efficacious

A variety of techniques such as slow cortical potentials, hemoencephalographic feedback, and cranial electrotherapy for treatment of ADHD have recently been reported. However, the majority of biofeedback studies have utilized EEG biofeedback; therefore, this technique will be the only one used to evaluate the efficacy for this disorder. The other techniques will be briefly presented at the end of this section. Even studies using EEG biofeedback to treat ADHD are difficult to summarize because they use a variety of training protocols and a variety of outcome measures. However, because the majority of studies used protocols that were directed toward reducing the abundance of slow frequencies while increasing the abundance of fast frequencies, some generalizations across studies are warranted. Numerous case studies; a multitude of treatment-only studies; some treatment compared to wait-list or no-treatment controls; and a few random-assignment, treatment-comparison groups have been reported. There are also a few review articles. These review articles should be evaluated with caution as they tend to have many of the same studies incorporated within their results. While the majority of the review articles conclude EEG biofeedback is effective when compared to no treatment, a placebo, or another treatment group, some of the reviews find fault with either the methodologies or outcome measurements of some studies.

Earlier uncontrolled studies using neurofeedback (NF) contingent on decreasing slow wave activity and increasing fast wave activity show persons with ADHD improved in symptoms, intelligence score, and academic performance (Grin'-Yatsenko et al. 2001; Lubar, Swartwood, Swartwood, & O'Donnell, 1995; Thompson & Thompson, 1998). In one study, only those individuals who significantly reduced theta over the training sessions showed a 12-point increase in Wisconsin Intelligent scale for Disorders Evaluation Scale (ADDES) rating score (Lubar et al. 1995). One large multicenter study (1,089 participants, aged five to 67 years) showed sensorimotor-beta EEG biofeedback training led to significant improvement in attentiveness, impulse control, and response variability as measured on the TOVA (Kaiser & Othmer, 2000) in those with moderate pretraining deficits.

A few early controlled studies compared EEG biofeedback to other treatments. The first of these was a study with four hyperkinetic children under six conditions: 1) no drug, 2) drug only, 3) drug and sensory motor rhythm (SMR) training, 4) drug and SMR reversal training, 5) drug and SMR training II, and 6) no drug and SMR training (Shouse & Lubar, 1979). Combining medication and SMR training resulted in substantial improvements in behavioral indices that exceeded the effects of drugs alone and were sustained with SMR training after medication was withdrawn. These changes were absent in the one highly distractible child who failed to acquire the SMR task.

In a study of 16 elementary-age children who were randomly assigned to conditions comparing EEG biofeedback to a waiting-list control, Carmody, Radvanski, Wadhvani, Sabo, and Vergara (2001) reported conflicting outcomes as measured by the TOVA and teacher reports. They found improvements in the reduction of errors of commission, anticipation, and attention, but no improvements in impulsivity or hyperactivity. Another small (n=18) controlled study showed increased intelligence scores and reduced inattentive behaviors as rated by parents in comparison to the waiting-list control (Linden, Habib, & Radojevic, 1996). Another study by Rossiter and La Vaque (1995) comparing EEG biofeedback to stimulant medication demonstrated both groups improved on measures of inattention, impulsivity, information processing, and variability as measured by the TOVA. Since 2002, a number of studies on the effectiveness of EEG biofeedback have been published, and they are presented briefly below. Some are outcome studies, and where available, the methodologies and outcome measures are presented while

others are reviews. Some studies were not based on slow-wave reduction and fast-wave enhancement, so their techniques need to be considered separately from the typical EEG biofeedback protocol. In a study of EEG biofeedback and stimulant medication effects, Fuchs, Birbaumer, Lutzenberger, Gruzelier, and Kaiser (2003) compared the effects of a three-month EEG biofeedback program providing reinforcement contingent on the production of cortical SMR (12-15 Hz) and beta-1 activity (15-18 Hz) with stimulant medication. Participants were aged eight to 12 years; 22 were assigned to the EEG biofeedback group and 12 to the methylphenidate group according to their parents' preference. Both EEG biofeedback and methylphenidate were associated with improvements on all subscales of the TOVA and on the speed and accuracy measures of the d2 Attention Endurance Test. Furthermore, behaviors related to the disorder were rated as significantly reduced in both groups by both teachers and parents on the IOWA-Connors Behavior Rating Scale. Another study relating stimulant medication to EEG biofeedback training reported 16 of 24 patients taking medications were able to lower their dose or discontinue medication totally after 30 sessions of EEG biofeedback (Alhambra, Fowler, & Alhambra, 1995). Finally, Monastra, Monastra, and George (2002) studied one hundred children with ADHD receiving Ritalin, parent counseling, and academic support at school. Based on parent preference, 50 children also received EEG biofeedback. While children improved on the TOVA and an ADHD evaluation scale while taking Ritalin, only those who had EEG biofeedback sustained these improvements without Ritalin. In a multiple case study (n=7), five participants completed an ABAB reversal methodology designed to alter the SMR/theta ratio in ADHD children (Heywood & Beale, 2003). Two participants failed to complete all training sessions, and the effects of training on behavior were analyzed both including and excluding these noncompleters. During alternate periods, they were trained using a placebo protocol identical to the treatment protocol except the association between EEG patterns and feedback was random. When all participants were included in analyses that controlled for overall trend, EEG biofeedback was found to be no more effective than the placebo control condition involving noncontingent feedback, and neither procedure resulted in improvements relative to baseline

levels. The such as maturation, history, and treatment order, but it does not control for carry-over from a treatment that has sustained effects, which EEG biofeedback has been shown to have in numerous studies. Because of a small number in the control group (n=2), possible carry-over effects, and a limited number of treatments (eight to 11), the reported lack of difference is tenuous at best. Prymachuk (2003) presented a review of randomized controlled trials (RCTs) evaluating treatment for 12 weeks in children with ADHD. Articles were selected if they were full reports published in any language in peer-reviewed journals. Fourteen RCTs (1,379 participants, 42% in one RCT) met the selection criteria. The findings relevant to EEG biofeedback state EEG biofeedback was superior to no treatment (one RCT), and treatment with EEG biofeedback led to better results on an intelligence test than did a waiting-list control (one RCT). In a replication of a previous study (Rossiter & La Vaque, 1995), Rossiter (2004) reports on a study with a larger sample, expanded age range, and improved statistical analysis. Thirty-one ADHD patients who chose stimulant drug treatment were matched with 31 patients who chose an EEG biofeedback treatment program. EEG biofeedback patients received either office (n = 14) or home (n = 17) EEG biofeedback. Stimulants for medication patients were titrated using the (TOVA). Both groups showed statistically and clinically significant improvement on the TOVA measures of attention, impulse control, processing speed, and variability in attention. The EEG biofeedback group demonstrated statistically and clinically significant improvement on behavioral measures (Behavior Assessment System for Children and Brown Attention Deficit Disorder Scales). The TOVA Confidence interval and nonequivalence null hypothesis testing confirmed the EEG biofeedback program produced outcomes equivalent to those obtained with stimulant drugs. To explore the effectiveness of EEG biofeedback on children with ADHD, a randomized selfcontrolled study with assessment taken before and after treatment was conducted (Chen et al. 2004). A total of 30 ADHD children were selected for the study from the Children's Mental Health Clinic of Nanjing Brain Hospital. Children were treated with EEG biofeedback. The Integrated Visual and Auditory continuous performance test (IVA) was used to evaluate before treatment and after 20 and 40 treatments. Main outcome measures were the control quotient and attention quotient of the IVA. After 20 treatments, the control quotients significantly increased and continued to significantly increase after 40

treatments. Cho et al. (2004) reported a study on the effectiveness of EEG biofeedback, along with virtual reality (VR), in reducing the level of inattention and impulsiveness. Twenty-eight male adolescents with social problems took part in this study. They were separated into three groups: a control group, a VR group, and a nonVR group. Both the VR and nonVR groups underwent eight sessions of EEG biofeedback training while the control group just waited during the same period. All participants performed a continuous performance task (CPT) before and after the complete training session. The results showed both the VR and nonVR groups (both also received EEG biofeedback training) achieved better scores in the CPT after training while the control group showed no significant difference. Eisenberg, Ben-Daniel, Mei-Tal, and Wertman (2004) reported a study to determine the effect of a new noninvasive technique of noncognitive biofeedback called Autonomic Nervous System Biofeedback Modality on the behavioral and attention parameters of a sample of children with attention deficit hyperactivity disorder. Nineteen subjects who met DSM-IV criteria for ADHD received four sessions of Autonomic Nervous System Biofeedback Modality treatment. The heart rate variability was measured before and after the treatment, as were measures of efficacy, including Connors Teacher Questionnaires (28 items), the Child Behavior Check List for parents and teachers, and Continuous Performance Test. Positive treatment effect was observed in all the subjects. A positive correlation between heart rate variability changes and improvement of symptoms of attention deficit hyperactivity disorder was found. learning problems for EEG biofeedback. Pre- and post-test reading and cognitive assessments were administered to sixth-, seventh-, and eighth-graders. Control and experimental groups were chosen at random. EEG biofeedback training was provided to the participants of the experimental group only. The control group had no treatment, just normal school-related activities. Seventeen students were assigned to each group. For various reasons, 12 finished treatment, and 14 were available for post measures in the control group. EEG biofeedback training lasted approximately 30 to 45 minutes and was conducted weekly for seven months. Some students received more sessions than others because of absences, field trips, testing, and other natural rhythms of home and school life. The average number of sessions

per student was 28. EEG biofeedback was significantly more effective in improving scores on reading tests than no EEG biofeedback training. There were significant interactions between EEG biofeedback and time on basic reading, and EEG biofeedback training was more effective in improving both the verbal and full-scale IQ scores than no EEG biofeedback training. There was a significant interaction between EEG biofeedback and time on verbal IQ and on full-scale IQ. There was a trend interaction for EEG biofeedback and performance IQ, but it was not significant. The results support the hypothesis that biofeedback training is effective in improving reading quotients and IQ in LD children.

In a study by Hanslmayr, Sauseng, Doppelmayr, Schabus, and Klimesch (2005), increasing upper alpha power while lowering theta in eight sessions improved cognitive functioning as measured by a mental rotation task performed before and after training. Only those subjects who were able to increase their upper alpha power performed better. Training success (extent of EEG biofeedback training-induced increase in upper alpha power) was positively correlated with the improvement in cognitive performance and significant increase in reference upper alpha power.

Fleischman and Othmer (2005) reported a case study of mildly developmentally delayed twins. They observed improvements in IQ scores and maintenance of the gains following EEG biofeedback. Full-scale IQ scores increased 22 and 23 points after treatment and were maintained at three follow-up retests over a 52-month period. ADHD symptom checklists completed by their mother showed a similar pattern of improvement and maintenance of gains.

Jacobs (2005) describes the application of EEG biofeedback with two children who manifested multiple diagnoses, including learning disabilities (LD), ADHD, social deficits, mood disorders, and pervasive developmental disorder (PDD). Both boys had adjusted poorly to school, family, and peers. They received individualized protocols based on their symptoms and functional impairments. They were administered semiweekly 20-minute sessions of one-channel EEG biofeedback training for approximately six months. In both cases, symptoms were identified and tracked with a parent rating scale and one case with the Symptom Assessment-45 questionnaire (SA-45) also. Each boy improved in all tracked symptoms without adverse effects.

In a study (Kropotov et al. 2005) of the effects of EEG biofeedback on Evoked Response Potentials (ERPs) in 86 ADHD children (ages nine to 14), ERPs were recorded in an auditory Go/No Go task before and after 15 to 22 sessions of EEG biofeedback. Each session consisted of 20 minutes of enhancing the ratio of the EEG power in the 15-18 Hz band compared to the EEG power in the rest of spectrum and seven to 10 minutes of enhancing the ratio of the EEG power in 12-15 Hz to the EEG power in the rest of spectrum. On the basis of quality of performance during training sessions, the patients were divided into two groups: good performers and bad performers. ERPs of good performers to Go and No Go cues gained positive components evoked within 180-420 ms latency. At the same time, no statistically significant differences between pre- and post-training ERPs were observed for bad performers. The ERP differences between post- and pre-treatment conditions for good performers were distributed over frontocentral areas and appear to reflect an activation of frontal cortical areas associated with beta training. A series of three studies by Li and colleagues are reported below: Li, Wu, & Chang, (2003) investigated the therapeutic effect of EEG biofeedback for ADHD. Sixty children aged six to 10 years were selected (30 children with attention deficit associated with hyperkinetic syndrome in the experimental group; 30 healthy children in the control group). The EEG recorded from the experiment group was significantly different from the control group. There was no significant difference in EEG between male and female children. Ten children received EEG biofeedback training and showed brain function was improved. In a second study by Li and Yu-Feng (2005), ADHD children with comorbid tic disorder (n=14) received EEG biofeedback treatment (average 34 sessions). The outcome was evaluated with a variety of outcome measures before and after treatment. Significant reductions in multiple symptoms were reported. Tic symptoms were greatly reduced in all but two children who also had Tourette's syndrome. In the third study (Li, Tang, et al. 2005), 113 outpatient children (88 male and 25 female, mean age of $10 \pm$ three years) from the Psychology Hyperactivity Department of the Central Hospital of Anshan City were selected. Inclusion criteria were from six to 14 years of age. Exclusion criteria were nervous system organic diseases, pervasive developmental disorder (PDD), mental retardation, epilepsy, psychotic disorder, and acoustical and visual abnormalities. ADHD

children were diagnosed, and then the EEG diagnostic accuracy was calculated. The diagnostic sensitivity of EEG on ADHD was 83.58%, the specificity was 82.61%, and misdiagnosis was 16.4%. These results compare favorably with the diagnostic accuracy of the Intermediate Visual and Auditory test (IVA). The EEG biofeedback system was also used for EEG biofeedback with 27 ADHD children. Conners Parent Symptom Questionnaire was used to assess pre- and post-hyperactivity levels. There was a significant difference between the EEG values before and after treatment, and the hyperactivity index scores were significantly declined from pre-treatment to post-treatment.

A study by Pop-Jordanova, Markovska-Simoska and Zorcec (2005) comprised 12 children of both sexes diagnosed as ADHD with the mean age of nine and a half years (seven to 13 years old). Each participated in a five-month program of EEG biofeedback training performed twice weekly. Posttreatment results showed improved EEG patterns expressed in increased 16-20 Hz (beta) activity and decreased 4-8 Hz (theta) activity. In parallel, higher scores on WISC-R, better school notes, and improved social adaptability and self-esteem were obtained.

A report by Putman, Othmer, Othmer, and Pollock (2005) that used the TOVA as the outcome measure was divided into three categories: a) primarily attentional deficits (n=12), b) primarily psychological complaints (n=20), and c) both (n=12). Participants were 44 males and females, six to 62 years old, who underwent treatment for a variety of clinical complaints. The TOVA was administered prior to EEG biofeedback training and 20 to 25 sessions thereafter. After EEG biofeedback training, significant improvements on omission, commission, and variability were observed. There was no change in reaction time. Reaction time was predominantly in the normal range for this population and remained unchanged following training.

Functional magnetic resonance imaging (fMRI) was used by Beauregard and Levesque (2006) to measure the effect of EEG biofeedback training in ADHD children. Twenty unmedicated ADHD children participated. Fifteen children were randomly assigned to the group trained to enhance the amplitude of the SMR (12-15 Hz) and beta 1 activity (15-18 Hz) and to decrease the amplitude of theta activity (4-7 Hz); whereas, the other five children were randomly assigned to the no-treatment group. Both groups were scanned one week before the beginning of EEG biofeedback and one week after the end of EEG

biofeedback while they performed a “Counting Stroop” task and a Go/No Go task. Changes were noted in several subcortical areas after biofeedback treatment in the EEG biofeedback group but not in the control group. These results suggest EEG biofeedback has the capacity to functionally normalize the brain systems mediating selective attention and response inhibition in ADHD children.

A study reported by Zhang, Zhang, and Jin (2006) compared EEG biofeedback with methylphenidate in ADHD children who were treated at the Department of Child Health Care, Xinhua Hospital. Participants were randomly assigned to groups. The EEG biofeedback group received treatments of reinforcing 16-20 Hz and suppressing 4-8 Hz; EEG biofeedback treatment was provided three to five times per week continuously for three months, totaling 35 to 40 sessions. The children in the medication group were treated with methylphenidate every morning. The dose started at 5 mg and increased gradually with the patients’ conditions until the effects were satisfied without any adverse effect. The Conners Parent Rating Scale was utilized to assess the behavioral changes. The children in the EEG biofeedback group and medication group were evaluated at pre-treatment, post-treatment and one, three, and six months of follow ups. Forty children who received EEG biofeedback and 16 who received medication were involved in the result analysis. Half the children who received EEG biofeedback were those who did not respond to medication after at least three months, so EEG biofeedback was provided.

After treatment, the EEG biofeedback group demonstrated significant decreases in scores on all factors of the Conners Parent Rating Scale compared to those at pretreatment and remained stable during a six month follow up. The medication group also showed significant decreases in scores of all factors except psychosomatic disorder and anxiety compared with those at pretreatment. The scores of psychosomatic disorder and anxiety were significantly lower in the EEG biofeedback group than in the medication group at post-treatment.

In a controlled study of effectiveness of EEG biofeedback training on children with ADHD, Zhong-Gui, Hai-Qing, and Shu-Hua (2006) reported EEG biofeedback training was applied for 30 minutes, two times per week for 40 sessions. The IVA was adopted to evaluate the effectiveness of EEG biofeedback training. The results from 60 children indicated the overall indexes of IVA were significantly

improved.

In a study by Kropotov et al. (2007), it was reported that changes in EEG spectrograms, event-related potentials, and event-related desynchronization were induced by relative beta training in ADHD children. EEG, ERPs, and event-related synchronization/desynchronization (ERD/ERS) were recorded and computed in an auditory Go/No Go task before and after 15 to 22 sessions of EEG biofeedback.

Eighty-six ADHD children participated in the study. Each session consisted of 30 minutes of relative beta training. The patients were divided into two groups (good performers and poor performers) depending on their ability to elevate beta activity during sessions. Amplitude of late positive components of evoked potentials in response to No Go stimuli increased, and event-related synchronization in alpha frequency band measured at central areas decreased in the group of good performers but did not change for the poor performers group. Evoked potential differences between post- and pre-treatment conditions for good performers were distributed over frontal-central areas, reflecting activation of frontal cortical areas associated with beta training. This activation likely indicates recovery of normal functioning of the executive system, but unfortunately, no clinical outcome measures were reported.

This study (Leins et al. 2007) compared EEG biofeedback training of theta-beta frequencies and training of slow cortical potentials (SCPs). SCP participants were trained to produce positive and negative

SCP shifts while the theta/beta participants were trained to suppress theta while increasing beta. Participants were blind to group assignment. Each group comprised 19 children with ADHD (aged eight to 13 years). Both groups were able to intentionally regulate cortical activity and improved in attention

and IQ. Parents and teachers reported significant behavioral and cognitive improvements.

Clinical effects

for both groups remained stable six months after treatment. Groups did not differ in behavioral or cognitive outcome.

A summary of recently published review articles is presented below. Most of the review articles include many of the same original studies; therefore, caution needs to be exercised in their interpretation.

Eighty-three studies were reviewed by Riccio and French (2004) to determine the status of treatments for ADHD. The studies were reviewed and categorized by the type of trial, whether or not the study included a control group, and the nature of the control group. The methodology of each study

was then rated and assigned to one of four categories (commendable, acceptable, marginal, and seriously flawed). The results were then categorized into three categories (positive, negative, and inconclusive).

Twenty studies were identified for treatment of ADHD with EEG biofeedback, and of those, seven were determined to have acceptable methodologies while 13 had marginal methodologies. The negative for one.

In another review, Fox, Tharp, and Fox (2005) reported that, in the last 30 years, multiple studies have consistently shown differences between ADHD children and nonADHD children in that the ADHD

children have a surplus of slow-wave activity, mostly in the delta and theta bands, and deficiencies in the

alpha and beta bands. They state that 70 to 80% of ADHD children respond favorably to stimulant

medication, 35% respond favorably to placebo, and 25 to 40% do not respond favorably to medication.

However, multiple studies have shown when stimulant medication is withdrawn, the improvements seen

during medication usage in the medication responders are no longer maintained. In a summary of five

EEG biofeedback outcome studies, they reported consistent improvements in behavior, IQ, and rating

scales comparable to medication usage, and only those trained in biofeedback maintained their improvements when the treatment was withdrawn.

In a review, Loo and Barkley (2005) report EEG measures have been used to study brain processes in children with ADHD for more than 30 years, and this research supports the EEG differences

between ADHD and nonADHD children. The differences are primarily in the frontal and central areas

with theta activity being more abundant and beta activity less abundant; therefore, the theta-beta ratio is

consistently and diagnostically larger in ADHD than nonADHD children. They report evidence of a

possible percentage of ADHD subtypes for which the EEG activity described above does not fit, and a

number of these individuals seem to be between 10 and 20% of all ADHD children. Thompson and

Thompson (2005) report these subtypes show distinctively different EEG patterns with an abundance of

high-frequency beta. The reviewers report that, more recently, EEG has been used, not only in research to

describe and quantify underlying neurophysiology of ADHD but also clinically in the assessment,

diagnosis, and treatment of ADHD. For the treatment of ADHD with EEG biofeedback, they

reported mixed results based on one study from an unpublished presentation at the American Psychological Association meeting in 1994 (so methodology and outcome assessment techniques cannot be determined) and three controlled studies. Of these three studies, one had a single-case design that was inappropriate for a treatment such as EEG biofeedback, which has a demonstrated carry-over effect. The two others demonstrated positive outcomes but were dismissed on what were viewed as weak methodical grounds because the studies did not use methodologies typically associated with pharmaceutical studies but used procedures usually associated with acceptable behavioral outcome studies.

In a series of review articles (Monastra, 2005; Monastra et al. 2005; Monastra et al. 2006), the authors report, in the past three decades, EEG biofeedback has emerged as a nonpharmacologic treatment for ADHD. These articles present imaging and EEG findings that support the theory of cortical hypoarousal, especially in the central and frontal regions of the cortex and that this intervention was derived from operant conditioning studies. These conditioning studies have demonstrated the capacity for neurophysiologic training in both humans and other mammals and targets atypical patterns of cortical activation that have been identified consistently in neuroimaging and quantitative EEG studies. The research findings published to date from case studies and controlled clinical outcome studies have reported increased cortical activation on quantitative electroencephalographic examination, improved attention and behavioral control, gains on tests of intelligence, improvement on self- and other rating scales, improved CPTs, and academic achievement. Three standard protocols of SMR enhancement and beta reduction, theta enhancement and beta reduction, and SMR enhancement and beta reduction are also presented.

A number of biofeedback articles based on techniques other than EEG biofeedback are presented below. These articles are presented in this section rather than the Emerging Applications section because they are treating individuals diagnosed with ADHD. The effect of ROSHI protocol and cranial electrotherapy stimulation on a nine-year-old anxious, dyslexic male with attention deficit disorder was studied by Overcash (2005). Psychological testing was administered, and QEEGs were recorded before and after treatment intervention. The patient was treated

using the ROSHI Complex Adaptive Protocol, Cranial Electrotherapy Stimulation, and the Project Read

Reading Program. This multimodal treatment lasted six months with follow-up testing administered 15

months after initial diagnostic testing. Before and after, objective psychological test results and QEEG

changes indicate significant improvement in reading, math, and spelling achievement and significant

reduction in anxiety and ADD symptoms.

Mize (2004) reported a single case study of hemoencephalography (HEG) with a 12-year-old male who had a well-established diagnosis of ADHD. He was performing well in school on

Concerta 36

mg at 7am and Ritalin 5 mg at 4pm. Off medication, he had significant abnormalities on IVA testing

(attention quotient or AQ = 78) and in the QEEG. IVA and clinical status measurements were made

before and after 10 sessions. Following the 10 sessions, the participant was tested off medication and

showed a normal QEEG with improved Z scores for relative power and a normal IVA (AQ = 99.75).

These results persisted in an 18-month follow up. His medication was lowered to Focalin 2.5 mg

twice

daily.

In a study designed to test the effectiveness of self-regulation of slow cortical potentials in children with ADHD (Strehl et al. 2006), 23 children with ADHD aged between eight and 13

years

received 30 sessions of self-regulation training of slow cortical potentials in three phases of 10 sessions

each. Feedback was provided while increasing and decreasing slow cortical potentials at central brain

regions. Measurement before and after the trials showed that children with ADHD learned to

regulate

negative slow cortical potentials. After training, significant improvement in behavior, attention, and IQ

score were observed. All changes proved to be stable at six months' follow up after the end of

training.

Clinical outcome was predicted by the ability to produce negative potential shifts in transfer sessions

without feedback. In summary, based on these studies and the reviews, EEG biofeedback has

typically

been shown to be superior to control conditions and equivalent to other treatments such as stimulant

medication.

The utilization of EEG measures to facilitate diagnostic determination and protocol determination

is strongly supported. Because the EEG protocols vary widely in specific bandwidths and thresholds selection, it is prudent for the practitioner to know the literature to determine which specific settings to use for each client. In addition to the EEG assessment, multiple assessments, including psychological, family, and medical history; a clinical interview; and standardized assessments, such as a continuous performance test and ratings scales, should be used to formulate a comprehensive treatment plan. EEG biofeedback techniques other than those focused on EEG patterns are also under development. Further studies are needed to examine long-term effects of training sessions and whether or not refresher sessions are needed to maintain the effects.

References

- Alhambra, M.A., Fowler, T.P., & Alhambra, A.A. (1995). EEG biofeedback: A new treatment option for ADD/ADHD. *Journal of Neurotherapy, 1*(2), 39-43.
- Beauregard, M., & Levesque, J. (2006). Functional magnetic resonance imaging investigation of the effects of EEG biofeedback training on the neural bases of selective attention and response inhibition in children with attention-deficit/hyperactivity disorder. *Applied Psychophysiology and Biofeedback, 31*(1), 3-20.
- Carmody, D.P., Radvanski, D.C., Wadhvani, S., Sabo, M.J., & Vergara, L. (2001). EEG biofeedback training and attention-deficit/hyperactivity disorder in an elementary school setting. *Journal of Neurotherapy, 4*(3), 5-27.
- Chen, Y., Jiao, G., Wang, C., Ke, X., Wang, M., & Chen, Y. (2004). Therapeutic effectiveness of electroencephalography biofeedback on children with attention deficit hyperactivity disorder. *Chinese Journal of Clinical Rehabilitation, 8*(18), 3690-3691.
- Cho, B.H., Kim, S., Shin, D.I., Lee, J.H., Lee, S.M., Kim, I.Y., et al. (2004). EEG biofeedback training with virtual reality for inattention and impulsiveness. *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society, 7*(5), 519-526.
- Eisenberg, J., Ben-Daniel, N., Mei-Tal, G., & Wertman, E. (2004). An autonomic nervous system biofeedback modality for the treatment of attention deficit hyperactivity disorder — an open pilot study. *The Israel Journal of Psychiatry and Related Sciences, 41*(1), 45-53.

Fleischman, M.J., & Othmer, S. (2005). Case study: Improvements in IQ score and maintenance of gains following EEG biofeedback with mildly developmentally delayed twins. *Journal of Neurotherapy*, 9(4), 35-46.

Fox, D.J., Tharp, D.F., & Fox, L.C. (2005). NF: An alternative and efficacious treatment for attention deficit hyperactivity disorder. *Applied Psychophysiology and Biofeedback*, 30(4), 365-373.

Fuchs, T., Birbaumer, N., Lutzenberger, W., Gruzelier, J.H., & Kaiser, J. (2003). Neurofeedback treatment for attention-deficit/hyperactivity disorder in children: A comparison with methylphenidate. *Applied Psychophysiology and Biofeedback*, 28(1), 1-12.

Grin'-Yatsenko, V.A., Kropotov, Yu., D., Ponomarev, V.A., Chutko, L.S., & Yakovenko, E.A. (2001). Effect of biofeedback training of sensorimotor and beta-sub-1EEG rhythms on attention parameters. *Human Physiology*, 27(3), 259-266.

Hanslmayr, S., Sauseng, P., Doppelmayr, M., Schabus, M., & Klimesch, W. (2005). Increasing individual upper alpha power by EEG biofeedback improves cognitive performance in human subjects. *Applied Psychophysiology and Biofeedback*, 30(1), 1-10.

Heywood, C., & Beale, I. (2003). EEG biofeedback vs. placebo treatment for attention deficit/hyperactivity disorder: A pilot study. *Journal of Attention Disorders*, 7(1), 43-55.

Jacobs, E.H. (2005). EEG biofeedback treatment of two children with learning, attention, mood, social, and developmental deficits. *Journal of Neurotherapy*, 9(4), 55-70.

Kaiser, D.A., & Othmer, S. (2000). Effect of neurofeedback on variables of attention in a large multicenter trial. *Journal of Neurotherapy*, 4(1), 5-15.

Kropotov, J.D., Grin'-Yatsenko, V.A., Ponomarev, V.A., Chutko, L.S., Yakovenko, E.A., & Nikishina, I.S. (2005). ERPs correlates of EEG relative beta training in ADHD children. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, 55(1), 23-34.

Kropotov, J.D., Grin'-Yatsenko, V.A., Ponomarev, V.A., Chutko, L.S., Yakovenko, E.A., & Nikishina, I.S. (2007). Changes in EEG spectrograms, event-related potentials, and event-related desynchronization induced by relative beta training in ADHD children. *Journal of Neurotherapy*, 11(2), 3-11.

Leins, U., Goth, G., Hinterberger, T., Klinger, C., Rumpf, N., & Strehl, U. (2007). EEG biofeedback for children with ADHD: A comparison of SCP and theta/beta protocols. *Applied Psychophysiology and Biofeedback*, 32(2), 73-88.

- Li, G., Wu, B., & Chang, S. (2003). Diagnosis and treatment for child attention deficit and hyperactivity disorder by biofeedback electroencephalograph. *Chinese Journal of Clinical Rehabilitation*, 7(22), 3104-3105.
- Li, L., & Yu-Feng, W. (2005). EEG biofeedback treatment on ADHD children with comorbid tic disorder. *Chinese Mental Health Journal*, 19(4), 262-265.
- Li, Y., Tang, Y., Liu, B., Long, S., Sun, G., Shen, L., et al. (2005). Electroencephalogram diagnosis and biofeedback treatment for the child with attention deficit hyperactivity disorder. *Chinese Journal of Clinical Rehabilitation*, 9(8), 236-237.
- Linden, M., Habib, T., & Radojevic, V. (1996). A controlled study of the effects of EEG biofeedback on cognition and behavior of children with attention deficit disorder and learning disabilities. *Biofeedback and Self Regulation*, 21(1), 35-49.
- Loo, S.K., & Barkley, R.A. (2005). Clinical utility of EEG in attention deficit hyperactivity disorder. *Applied Neuropsychology*, 12(2), 64-76.
- Lubar, J.F., Swartwood, M.O., Swartwood, J.N., & O'Donnell, P.H. (1995). Evaluation of the effectiveness of EEG neurofeedback training for ADHD in a clinical setting as measured by changes in TOVA scores, behavioral ratings, and WISC-R performance. *Biofeedback and Self Regulation*, 20(1), 83-99.
- Mize, W. (2004). Hemoencephalography — A new therapy for attention deficit hyperactivity disorder (ADHD): Case report. *Journal of Neurotherapy*, 8(3), 77-97.
- Monastra, V.J. (2005). Electroencephalographic biofeedback (neurotherapy) as a treatment for attention deficit hyperactivity disorder: Rationale and empirical foundation. *Child and Adolescent Psychiatric Clinics of North America*, 14(1), 55-82.
- Monastra, V.J., Lynn, S., Linden, M., Lubar, J.F., Gruzelier, J., & La Vaque, T.J. (2006). Electroencephalographic biofeedback in the treatment of attention-deficit/hyperactivity disorder. *Journal of Neurotherapy*, 9(4), 5-34.
- Monastra, V.J., Lynn, S., Linden, M., Lubar, J.F., Gruzelier, J., & LaVaque, T.J. (2005). Electroencephalographic biofeedback in the treatment of attention-deficit/hyperactivity disorder. *Applied Psychophysiology and Biofeedback*, 30(2), 95-114.
- Monastra, V.J., Monastra, D.M., & George, S. (2002). The effects of stimulant therapy, EEG biofeedback, and parenting style on the primary symptoms of attention-deficit/hyperactivity disorder. *Applied Psychophysiology and Biofeedback*, 27(4), 231-249.
- Orlando, P.C., & Rivera, R.O. (2004). Elementary students with identified learning problems.

- Journal of Neurotherapy*, 8(2), 5-19.
- Overcash, S.J. (2005). The effect of ROSHI protocol and cranial electrotherapy stimulation on a nine-year-old anxious, dyslexic male with attention deficit disorder: A case study. *Journal of Neurotherapy*, 9(2), 63-77.
- Pop-Jordanova, N., Markovska-Simoska, S., & Zorcec, T. (2005). EEG biofeedback treatment of children with attention deficit hyperactivity disorder. *Prilozi/Makedonska akademija na naukite i umetnostite, Oddelenie za biologski i medicinski nauki — Contributions/Macedonian Academy of Sciences and Arts, Section of Biological and Medical Sciences*, 26(1), 71-80.
- Prymachuk, S. (2003). Review: Extended stimulant medication is effective in children with attention deficit hyperactivity disorder. *Evidence-Based Nursing*, 6(1), 11-11.
- Putman, J.A., Othmer, S.F., Othmer, S., & Pollock, V.E. (2005). TOVA results following interhemispheric bipolar EEG training. *Journal of Neurotherapy*, 9(1), 37-52.
- Riccio, C.A., & French, C.L. (2004). The status of empirical support for treatments of attention deficits. *The Clinical Neuropsychologist*, 18(4), 528-558.
- Rossiter, T.R. (1998). Patient-directed neurofeedback for ADHD. *Journal of Neurotherapy*, 2(4), 54-63.
- Rossiter, T.R., & La Vaque, T.J. (1995). A comparison of EEG biofeedback and psychostimulants in treating attention deficit/hyperactivity disorders. *Journal of Neurotherapy*, 1(1), 48-59.
- Rossiter, T. (2004). The effectiveness of EEG biofeedback and stimulant drugs in treating AD/HD: Part II. Replication. *Applied Psychophysiology and Biofeedback*, 29(4), 233-243.
- Shouse, M.N., & Lubar, J.F. (1979). Operant conditioning of EEG rhythms and Ritalin in the treatment of hyperkinesis. *Biofeedback and Self Regulation*, 4(4), 299-312.
- Strehl, U., Leins, U., Goth, G., Klinger, C., Hinterberger, T., & Birbaumer, N. (2006). Self-regulation of slow cortical potentials: A new treatment for children with attention-deficit/hyperactivity disorder. *Pediatrics*, 118(5), e1530-1540.
- Thompson, L., & Thompson, M. (1998). Neurofeedback combined with training in metacognitive strategies: Effectiveness in students with ADD. *Applied Psychophysiology and Biofeedback*, 23(4), 243-263.
- Thompson, L., & Thompson, M. (2005). Neurofeedback intervention for adults with ADHD. *Journal of*

Adult Development, 12(2), 123-130.

Zhang, F., Zhang, J., & Jin, X. (2006). Effect of electroencephalogram biofeedback on behavioral problems of children with attention deficit hyperactivity disorder. *Chinese Journal of Clinical Rehabilitation*, 10(10), 74-76.

Zhong-Gui, X., Hai-Qing, X., & Shu-Hua, S. (2006). The controlled study of effectiveness of EEC biofeedback training on children with attention deficit hyperactivity disorder. *Chinese Journal of Clinical Psychology*, 14(2), 207-208.

4. ANXIETY

Much of the information provided here is from Carolyn Yucca's 2008 book "Evidence Based Practice in Biofeedback & Neurofeedback" AAPB, Wheat Ridge, CO.

Overview & Efficacy: Everybody gets anxious. Treatment is called for if the amount of anxiety is out of proportion to the problem or lasts too long. Many methods for helping people reduce and control their anxiety have been shown to be effective. Behavioral techniques include relaxation training, cognitive restructuring, and biofeedback. Any form of biofeedback which helps people become aware of their physiological responses as they become anxious and which helps people learn to relax is apparently at least as effective as any other behavioral technique.

This therapy is rated as efficacious (level 4 on a scale of 1 – 5 with 5 being the best).

Why biofeedback would help this problem: There are several different underlying problems which cause abnormal levels of anxiety. Biofeedback helps each for different reasons.

a. Breathing problems which cause anxiety: Half or more of people who habitually breathe too rapidly with shallow breaths are anxious because of the effects of their breathing on their brains' chemistry. Most of these people are not aware they have incorrect breathing patterns. These incorrect patterns are easily detected using psychophysiological assessments and are corrected using several types of biofeedback related to helping people normalize their breathing patterns. When the breathing is normalized, the anxiety goes away.

b. When a person experiences greater levels of anxiety or the anxiety lingers far longer than it should, the body's normal responses to an emergency situation don't shut down. This can cause the body to wear out while thinking and memory patterns change. The physiological reactions to anxiety are accurately assessed using psychophysiological recording techniques so both the patient and therapist always know when any therapy is helping and how much. Biofeedback treatments show the patient the abnormal physiological response levels. Patients use this knowledge to recognize when they are becoming abnormally anxious and to control their anxiety.

Brief summary of evidence supporting the efficacy of biofeedback for abnormal levels of anxiety: Anxiety

Level 4: Efficacious

Multiple case studies have demonstrated clinically significant outcomes with carefully screened and thoroughly assessed participants for various forms of anxiety-related disorders. There are also several treatment-only group studies with moderate sample sizes, demonstrating positive results of various forms of biofeedback that were often combined with other behavioral interventions. A few well-controlled, randomized studies have shown biofeedback to be equivalent to other relaxation and self-control methods for reducing anxiety while it is occasionally shown to be superior to another intervention. Most show biofeedback (EMG, GSR, thermal, or neurofeedback) to be roughly equivalent to progressive relaxation or meditation.

Lehrer, Carr, Sargunraj, and Woolfolk (1994) evaluated the hypothesis that biofeedback is most effective when applied in the same modality as the disorder (autonomic feedback for ANS disorders, EMG feedback for muscular disorders, etc.). Other researchers have asserted self-relaxation techniques have in common the process of using conscious intent to calm oneself, and for anxiety reduction, it may matter little which modality is used because the central component is the cognitively based conscious intent. Clarification of this issue must await further clinical outcome studies.

Two studies showed biofeedback's efficacy in reducing anxiety without making comparisons with other relaxation techniques. Hurley and Meminger (1992) used frontal EMG biofeedback with 40 subjects trained to criterion and assessed anxiety over time using the State-Trait Anxiety Inventory (STAI). State anxiety improved more than trait anxiety. Wenck, Leu, and D'Amato (1996) trained 150 seventh- and eighth-graders with thermal and EMG feedback and found significant reduction in state and trait anxiety.

Roome and Romney (1985) compared progressive muscle relaxation to EMG biofeedback training with 30 children and found an advantage for biofeedback; however, Scandrett, Bean, Breeden, and Powell (1986) found some advantage of progressive muscle relaxation over EMG biofeedback in reducing anxiety in adult psychiatric inpatients and outpatients.

Rice, Blanchard, and Purcell (1993) studied reduction in generalized anxiety by comparing groups given EMG frontal feedback, EEG alpha-increase feedback, and EEG alpha-decrease feedback to two control conditions (a pseudo-meditation condition and a wait-list control). All treatment

groups had comparable and significant decreases in the STAI and drops in the Psychosomatic Symptom Checklist. The alpha-increasing biofeedback condition produced one effect not found with the other treatment conditions: a reduction in heart-rate reactivity to stressors. Similar results were obtained by Sarkar, Rathee, and Neera (1999), who compared the generalized anxiety disorder response to pharmacotherapy and to biofeedback; the two treatments had similar effects on symptom reduction. Hawkins, Doell, Lindseth, Jeffers, and Skaggs (1980) concluded, from a study with 40 hospitalized schizophrenics, that thermal biofeedback and relaxation instructions had an equivalent effect on anxiety reduction. However, Fehring (1983) found adding GSR biofeedback to a Benson-type relaxation technique reduced anxiety symptoms more than relaxation alone. Vanathy, Sharma, and Kumar (1998), applying EEG biofeedback to generalized anxiety disorder, compared increased alpha with increased theta. The two procedures were both effective in decreasing symptoms. In a recent case study, Hammond (2003) reported on two cases using EEG biofeedback for OCD. Clinically significant improvements for both participants were reported. In a single case study (Goodwin & Montgomery, 2006) of a 39-year-old male with panic disorder and agoraphobia, electrodermal biofeedback was combined with CBT, graded exposure. They reported a complete cessation of panic attacks, a remission of agoraphobia, and a clinically significant reduction in depression. In a study by Gordon, Staples, Blyta, and Bytyqi (2004) a total of 139 PTSD postwar high school students were provided a six-week program of biofeedback, meditation, drawings, autogenics, guided imagery, genograms, and breathing techniques. No control group was used, but they reported a significant reduction immediately after treatment and at follow up. In a two-treatment group comparison study (n=50) of anxiety in individuals with chronic pain, Corrado, Gottlieb, and Abdelhamid (2003) reported a significant improvement in anxiety and somatic complaints in the group that received biofeedback of finger temperature increase and muscle tension reduction when compared to a pain education group. In an RCT study of 87 participants, Bont, Castilla, and Maranon (2004) presented the outcome of three intervention programs applied to fear of flying: a reattributorial training-based program, a

mixedexposure

procedure, and finally a biofeedback training program in order to change psychophysiological responses. A fourth group of wait-list controls were also assessed. They found a significant reduction in

anxiety for the treatment groups when compared to the control group of no treatment. In another RCT

study of imipramine and imipramine plus biofeedback, Coy, Cardenas, Cabrera, Zirot, and Claros (2005)

found the biofeedback group plus medication (n=18) was significantly improved compared to the medication-only group (n=14).

From a group of 312 high school students in Shanghai, Dong and Bao (2005) recruited 70 students who met criteria for high levels of anxiety and assigned 35 students to a group who were treated

with biofeedback and 35 to a group of no-treatment controls. They reported a significant improvement in

anxiety, somatization, and depression in the treatment group when compared to the controls.

In conclusion, biofeedback of various modalities is effective for anxiety reduction. It is often found to compare favorably with other behavioral techniques and occasionally found to be superior to

those and medication alone.

References

Bont, J.I.C., Castilla, C.D.S., & Maranon, P.P. (2004). Comparison of three fear of flying therapeutic programs. *Psicothema*, 16(4), 661-666.

Corrado, P., Gottlieb, H., & Abdelhamid, M.H. (2003). The effect of biofeedback and relaxation training

on anxiety and somatic complaints in chronic pain patients. *American Journal of Pain Management*, 13(4), 133-139.

Coy, P.C., Cardenas, S.J., Cabrera, D.M., Zirot, G.Z., & Claros, M.S. (2005).

Psychophysiological and

behavioral treatment of anxiety disorder. *Salud Mental*, 28(1), 28-37.

Dong, W., & Bao, F. (2005). Effects of biofeedback therapy on the intervention of examination-caused

anxiety. *Chinese Journal of Clinical Rehabilitation*, 9(32), 17-19.

Fehring, R.J. (1983). Effects of biofeedback-aided relaxation on the psychological stress symptoms of

college students. *Nursing Research*, 32(6), 362-366.

Goodwin, E.A., & Montgomery, D.D. (2006). A cognitive-behavioral, biofeedback-assisted relaxation

treatment for panic disorder with agoraphobia. *Clinical Case Studies*, 5(2), 112-125.

Gordon, J.S., Staples, J.K., Blyta, A., & Bytyqi, M. (2004). Treatment of posttraumatic stress disorder in

postwar Kosovo high school students using mind-body skills groups: A pilot study. *Journal of Traumatic*

Stress, 17(2), 143-147.

Hammond, D.C. (2003). QEEG-guided neurofeedback in the treatment of obsessive-compulsive disorder.

Journal of Neurotherapy, 7(2), 25-52.

Hawkins, R.C., II, Doell, S.R., Lindseth, P., Jeffers, V., & Skaggs, S. (1980). Anxiety reduction in

hospitalized schizophrenics through thermal biofeedback and relaxation training. *Perceptual &*

Motor Skills, 51(2), 475-482. Hurley, J.D., & Meminger, S.R. (1992). A relapse-prevention

program: Effects of electromyographic

training on high and low levels of state and trait anxiety. *Perceptual and Motor Skills*, 74(3, Pt.

1),

699-705.

Lehrer, P.M., Carr, R., Sargunraj, D., & Woolfolk, R.L. (1994). Stress management techniques:

Are they

all equivalent, or do they have specific effects? *Biofeedback & Self Regulation*, 19(4), 353-401.

Rice, K.M., Blanchard, E.B., & Purcell, M. (1993). Biofeedback treatments of generalized

anxiety

disorder: Preliminary results. *Biofeedback & Self-Regulation*, 18(2), 93-105.

Roome, J.R., & Romney, D.M. (1985). Reducing anxiety in gifted children by inducing

relaxation.

Roepers Review, 7(3), 177-179.

Sarkar, P., Rathee, S.P., & Neera, N. (1999). Comparative efficacy of pharmacotherapy and

biofeedback

among cases of generalised anxiety disorder. *Journal of Projective Psychology & Mental Health*,

6(1),

69-77.

Scandrett, S.L., Bean, J.L., Breeden, S., & Powell, S. (1986). A comparative study of

biofeedback and

progressive relaxation in anxious patients. *Issues in Mental Health Nursing*, 8(3), 255-271.

Vanathy, S., Sharma, P.S.V.N., & Kumar, K.B. (1998). The efficacy of alpha and theta

neurofeedback

training in treatment of generalized anxiety disorder. *Indian Journal of Clinical Psychology*,

25(2),

136-143.

Wenck, L.S., Leu, P.W., & D'Amato, R.C. (1996). Evaluating the efficacy of a biofeedback

intervention

to reduce children's anxiety. *Journal of Clinical Psychology*, 52(4), 469-473.

Respiration – Anxiety Link:

DeGuire S, Gevirtz R, Hawkinson D, Dixon K: Breathing retraining: a three-year follow-up study of treatment for hyperventilation syndrome and associated functional cardiac symptoms, Biofeedback Self Regul, 21:191-8, 1996.

Gevirtz R: Resonant frequency training to restore autonomic homeostasis for treatment of psychophysiological disorders. Biofeedback 27: 7-9, 1999.

Lehrer P, Vaschillo E, Vaschillo B: Resonant frequency biofeedback training to increase cardiac

variability: Rationale and manual for training. Applied psychophysiology and biofeedback 25:177 – 191, 2000.

5. Raynaud's Disease

Level 4: Efficacious

There were several brief, relatively uncontrolled studies published in the 1970s that confirmed the rationale underlying temperature biofeedback (TBF) treatment of primary Raynaud's disease (RP). Peterson and Vorhies (1983) studied thermal biofeedback-trained Raynaud's patients, observing the speed of hand temperature return to baseline after hand immersion in ice water, which was six to seven times as fast after biofeedback training (six minutes average after training versus 40 minutes before). Jobe, Sampson, Roberts, and Kelly (1986) compared hand temperature responses to whole-body chilling before and after biofeedback training and found it to be effective. When Guglielmi, Roberts, and Patterson (1982) compared thermal biofeedback with EMG biofeedback and controls with a double-blind procedure, all three groups had comparable improvements, suggesting a role of nonspecific factors. The results of this study have limited generalization to clinical practice because the participants could not have adequate instructions about how to perform the physiological changes, when and how to utilize the training, and any motivational guidelines for incorporating the training daily to enhance the clinical training. Keefe, Surwit, and Pilon (1980) found similar results, in which other behavioral control methods performed as well as thermal biofeedback. However, Freedman et al. (1988) compared simple thermal biofeedback with autogenic training and found the former to be more effective. The largest study to date of Raynaud's involving biofeedback compared use of a calcium-channel blocker (nifedipine) with thermal biofeedback, EMG feedback, and a placebo (Raynaud's Treatment Study Investigators, 2000). In this study of 313 subjects with primary Raynaud's disease, nifedipine seemed to be the superior agent for reducing symptoms. Problems with training the thermal biofeedback subjects to an adequate level of skill, however, mitigated the final results (Middaugh et al. 2001). A recent review of finger temperature training in primary Raynaud's phenomenon that focused on whether subjects were adequately trained to increase finger temperature found eight RCT, one

nonRCT, and two follow-up studies (Karavidas, Tsai, Yucha, McGrady, & Lehrer, 2006). The authors concluded the level of evidence for TBF efficacy is categorized as Level IV: efficacious. The rationale was based on three randomized controlled trials conducted in independent laboratories that demonstrated “superiority or equivalence” of treatments that include TBF.

References

Freedman, R.R., Sabharwal, S.C., Ianni, P., Desai, N., Wenig, P., & Mayes, M. (1988). Nonneural betaadrenergic vasodilating mechanism in temperature biofeedback. *Psychosomatic Medicine*, 50(4), 394-401.

Guglielmi, R.S., Roberts, A.H., & Patterson, R. (1982). Skin temperature biofeedback for Raynaud’s disease: A double-blind study. *Biofeedback & Self-Regulation*, 7(1), 99-120.

6. Urinary Incontinence in Females

Level 5: Efficacious and Specific

Numerous within-subject studies have demonstrated biofeedback efficacy at the lower levels of efficacy (Dannecker, Wolf, Raab, Hepp, & Anthuber, 2005; Rett et al. 2007); all of these have not been reported here. Rather, only RCTs and systematic reviews are included that show levels four and five efficacy of biofeedback for urinary incontinence in females. It is better than no treatment (i.e., control) (Burgio et al. 1998; Burns et al. 1993; Dougherty et al. 2002; McDowell et al. 1999), better than or equal to other behavioral treatments (e.g., pelvic floor exercises, bladder training) (Burns et al. 1993; Glavind, Nohr, & Walter, 1996; Sherman, Davis, & Wong, 1997; Sung, Hong, Choi, Baik, & Yoon, 2000; Weatherall, 1999; Wyman, Fantl, McClish, & Bump, 1998; Wallace, Roe, Williams, & Palmer, 2004), as effective as pelvic floor electrical stimulation (Goode et al. 2003; Wang, Wang, & Chen, 2004) and vaginal cone (Seo, Yoon, & Kim, 2004), and better than drug (i.e., oxybutynin chloride) treatment (Burgio et al. 1998; Goode, 2004). The benefit of biofeedback over drug therapy was supported by a systematic review (Teunissen, de Jonge, van Weel, & Lagro-Janssen, 2004). Combining drug and behavioral therapy in a stepped program can produce added benefit for those not satisfied with the outcome of single treatment (Burgio, Locher, & Goode, 2000). Biofeedback is also effective for reducing urinary incontinence in older women (Tadic et al.

2007). In comparison to drug treatment with oxybutynin, biofeedback reduced incontinence (Goode, 2004) and nocturia in older women (Johnson, Burgio, Redden, Wright, & Goode, 2005). Exploring the effect of pelvic floor muscle exercises on urinary incontinence following childbirth is more complicated. Studies where it is administered prenatally include women who are both continent and incontinent postnatally; this diminishes the results, and the effect is not different from that seen in control groups. However, in studies in which this training is provided to only those who are incontinent after childbirth, there is a significant effect on reducing or resolving urinary incontinence (Haddow, Watts, & Robertson, 2005). In those with multiple sclerosis, EMG biofeedback for lower urinary tract dysfunction, especially in combination with neuromuscular electrical stimulation, decreased incontinence episodes (McClurg, Ashe, & Lowe-Strong, 2007). A number of systematic reviews are now available reporting efficacy for pelvic floor muscle training (Bø, 2003; Neumann, Grimmer, & Deenadayalan, 2006; Hay-Smith & Dumoulin, 2006). In a Cochrane Review, Alhasso, McKinlay, Patrick, and Stewart (2006) found symptomatic improvement was more common among those on anticholinergic drugs compared with bladder training (with and without biofeedback). In contrast, a more specific review of pelvic floor muscle biofeedback reported the overall mean treatment improvement was 72.6% and that in 60% of paired comparisons, biofeedback demonstrated superior symptomatic outcome to control or alternate treatment groups, including oxubutynin (Glazer & Laine, 2006). Recent studies have explored variations in biofeedback therapy. Home biofeedback for 12 weeks resulted in an increase in pelvic floor muscle activity and a decrease in leakage index (Aukee et al. 2004). A telemedicine continence program (including biofeedback-assisted pelvic floor training) was as effective as a clinic-based program (Hui, Lee, & Woo, 2006). Position during training (supine vs supine and upright) does not differentially affect treatment outcomes (France, Zyczynski, Downey, Rause, & Wister, 2006).

References

Alhasso, A.A., McKinlay, J., Patrick, K., & Stewart, L. (2006). Anticholinergic drugs versus non-drug active therapies for overactive bladder syndrome in adults. *Cochrane Database of Systematic Reviews*

(Online), 4(4), CD003193.

Aukee, P., Immonen, P., Laaksonen, D.E., Laippala, P., Penttinen, J., & Airaksinen, O. (2004). The effect of home biofeedback training on stress incontinence. *Acta Obstetrica et Gynecologica Scandinavica*, 83(10), 973-977.

Bø, K. (2003). Is there still a place for physiotherapy in the treatment of female incontinence? *EAU Update Series*, 1(3), 145-153.

Burgio, K.L., Locher, J.L., & Goode, P.S. (2000). Combines behavioral and drug therapy for urge incontinence in older women. *Journal of the American Geriatric Society*, 48(4), 370-374.

Burgio, K.L., Locher, J.L., Goode, P.S., Hardin, J.M., McDowell, B.J., Dombrowski, M., et al. (1998). Behavioral vs. drug treatment for urge urinary incontinence in older women: A randomized controlled trial. *Journal of the American Medical Association*, 280(23), 1995-2000.

Burns, P.A., Pranikoff, K., Nochajski, T.H., Hadley, E.C., Levy, K.J., & Ory, M.G. (1993). A comparison of effectiveness of biofeedback and pelvic muscle exercise treatment of stress incontinence in older community-dwelling women. *Journal of Gerontology*, 48(4), M167-174.

Dannecker, C., Wolf, V., Raab, R., Hepp, H., & Anthuber, C. (2005). EMG-biofeedback assisted pelvic floor muscle training is an effective therapy of stress urinary or mixed incontinence: A seven-year experience with 390 patients. *Archives of Gynecology and Obstetrics*, 273(2), 93-97.

Dougherty, M.C., Dwyer, J.W., Pendergast, J.F., Boyington, A.R., Tomlinson, B.U., Coward, et al. (2002). A randomized trial of behavioral management for continence with older rural women. *Research in Nursing and Health*, 25(1), 3-13.

France, D.F., Zyczynski, H.M., Downey, P.A., Rause, C.R., & Wister, J.A. (2006). Effect of pelvic-floor muscle exercise position on continence and quality-of-life outcomes in women with stress urinary incontinence. *Physical Therapy*, 86(7), 974-986.

Glavind, K., Nohr, S.B., & Walter, S. (1996) Biofeedback and physiotherapy versus physiotherapy alone in the treatment of genuine stress urinary incontinence. *International Urogynecologic Journal of Pelvic Floor Dysfunction*, 7(6), 339-343.

Glazer, H.I., & Laine, C.D. (2006). Pelvic floor muscle biofeedback in the treatment of urinary incontinence: A literature review. *Applied Psychophysiology and Biofeedback*, 31(3), 187-201.

Goode, P.S., Burgio, K.L., Locher, J.L., Roth, D.L., Umlauf, M.G., Richter, H.E., et al. (2003). Effect of behavioral training with or without pelvic floor electrical stimulation on stress incontinence in

women: A randomized controlled trial. *JAMA*, 290(3), 345-352.

Goode, P.S. (2004). Behavioral and drug therapy for urinary incontinence. *Urology*, 63(3 Suppl. 1), 58-64.

Haddow, G., Watts, R., & Robertson, J. (2005). Effectiveness of a pelvic floor muscle exercise program on urinary incontinence following childbirth. *International Journal of Evidence-Based Healthcare*, 3(5), 103-146.

Hay-Smith, E.J., & Dumoulin, C. (2006). Pelvic floor muscle training versus no treatment, or inactive control treatments, for urinary incontinence in women. *Cochrane Database of Systematic Reviews (Online)*, 1(1), CD005654.

Hui, E., Lee, P.S., & Woo, J. (2006). Management of urinary incontinence in older women using videoconferencing versus conventional management: A randomized controlled trial. *Journal of Telemedicine and Telecare*, 12(7), 343-347.

Johnson, T.M.I., Burgio, K.L., Redden, D.T., Wright, K.C., & Goode, P.S. (2005). Effects of behavioral and drug therapy on nocturia in older incontinent women. *Journal of the American Geriatric Society*, 53(5), 846-850.

McClurg, D., Ashe, R.G., & Lowe-Strong, A.S. (2007). Neuromuscular electrical stimulation and the treatment of lower urinary tract dysfunction in multiple sclerosis: A double-blind, placebo-controlled, randomised clinical trial. *Neurourology and Urodynamics*, (Epub, ahead of print).

McDowell, B.J., Engberg, S., Sereika, S., Donovan, N., Jubeck, M.E., Weber, E., et al. (1999). Effectiveness of behavioral therapy to treat incontinence in homebound older adults. *Journal of the American Geriatric Society*, 47(3), 309-318.

Neumann, P.B., Grimmer, K.A., & Deenadayalan, Y. (2006). Pelvic floor muscle training and adjunctive therapies for the treatment of stress urinary incontinence in women: A systematic review. *BMC Women's Health*, 6, 11.

Rett, M.T., Simoes, J.A., Herrmann, V., Pinto, C.L., Marques, A.A., & Morais, S.S. (2007). Management of stress urinary incontinence with surface electromyography-assisted biofeedback in women of reproductive age. *Physical Therapy*, 87(2), 136-142.

Seo, J.T., Yoon, H., & Kim, Y.H. (2004). A randomized prospective study comparing new vaginal cone and FES-biofeedback. *Yonsei Medical Journal*, 45(5), 879-884.

Sherman, R.A., Davis, G.D., & Wong, M.F. (1997). Behavioral treatment of exercise-induced

urinary incontinence among female soldiers. *Military Medicine*, 162(10), 690-704.

Sung, M.S., Hong, J.Y., Choi, Y.H., Baik, S.H., Yoon, H. (2000). FES-biofeedback versus intensive pelvic floor muscle exercise for the prevention and treatment of genuine stress incontinence. *Journal of Korean Medical Science*, 15(3), 303-308.

Tadic, S.D., Zdaniuk, B., Griffiths, D., Rosenberg, L., Schafer, W., & Resnick, N.M. (2007). Effect of biofeedback on psychological burden and symptoms in older women with urge urinary incontinence. *Journal of the American Geriatric Society*, 55(12), 2010-2005.

7. Chronic Pain

Level 4: Efficacious

Chronic pain can arise from just one or two sites, or it can be pervasive and widespread. Most research studies focus on pain from a particular site, but because chronic pain, regardless of its source, may involve nonspecific factors such as neural sensitization, altered neurotransmitter levels, inflammation, and muscle guarding, there is some logic to also treating chronic pain as a unitary condition regardless of its site and supposed generating mechanism. This section on Chronic Pain excludes specific categories that are presented in other sections for that disorder (e.g., headaches). Because some specific disorders have clearly demonstrated biofeedback effectiveness while others have only case studies and mixed results for the efficacy of specific disorders, it is necessary to generalize across various specific pain disorders. For specific disorders, review other sections of this document and other related, more detailed publications such as AAPB's White Paper on chronic pain (Clinical Efficacy of Psychophysiological Assessments and Biofeedback Interventions for Chronic Pain Disorders Other Than Head-Area Pain, 2006). Most studies of biofeedback treatment are from studies where biofeedback is a part of a multiple modality program, so it is not possible at this time to ascertain the unique contributions biofeedback may provide for chronic pain patients. However, the studies presented below clearly demonstrate treatment programs that include biofeedback are as effective as standard (single treatment or medication alone) and more effective than no-control conditions. Flor and Birbaumer (1993) studied both EMG biofeedback and cognitive therapy for both back pain and temporomandibular joint pain. In this study, biofeedback had the strongest effect on many aspects of pain, and the effects were still present at a 24-month follow up. Vlaeyen, Haazen, Schuerman,

Kole-Snijders, and van Eek (1995) studied the response to EMG biofeedback training in 71 chronic back pain patients in comparison with a cognitive-training group. The groups had comparable positive outcomes as compared to wait-list control and an operant conditioning-only treatment. Newton-John, Spence, and Schotte (1995) compared cognitive therapy with EMG biofeedback in chronic back patients and obtained similar beneficial effects with both as compared to a wait-list control group. Effects persisted at a six-month follow up. Humphreys and Gevirtz (2000) reported a study of recurrent abdominal pain in 64 children and teenagers that used thermal biofeedback alone or in combination with cognitive-behavioral treatment. Results for pain relief were significantly above an inactive treatment (fiber-only) control group.

A comprehensive literature review of biopsychosocial approaches to chronic pain published in 2001 (Nielson & Weir, 2001) examined many single and combined treatments and found EMG biofeedback had at least moderate support as a separate treatment. The bulk of the studies and the three systematic reviews covered mostly back pain, the most common focus for research at that time. Fifty chronic pain patients were evaluated pre- and post-treatment using the Wahler Physical Symptoms Checklist and the IPAT Anxiety Scale (Corrado, Gottlieb, & Abdelhamid, 2003). Participants were randomly assigned to a biofeedback-plus-relaxation-training group or a pain-education group. The biofeedback-plus-relaxation-training group reported significantly improved symptoms of anxiety and significantly reduced somatic complaints in comparison with the pain-education group.

Hawkins and Hart (2003) used thermal biofeedback in the treatment of pain associated with endometriosis. A multiple case study design (n = 5) was employed. Four participants were able to demonstrate mastery over hand temperature through thermal biofeedback. Of those four participants, significant reductions in various aspects of pain were observed. Pulliam and Gatchel (2003) examined the literature with respect to biofeedback and chronic pain and summarized the current indications of this treatment modality for various disorders. Conditions reviewed included headaches, temporomandibular disorders, low back pain, fibromyalgia, irritable bowel syndrome, and Raynaud's disease. The authors concluded biofeedback represents a useful adjunctive treatment technique for most chronic pain conditions. Its addition to standard treatment provides significant incremental validity for many disorders.

A review article by Stinson (2003) reported only RCT trials comparing a clearly defined psychological treatment with a control condition (wait-list and self-monitoring) for chronic pain

in children or adolescents. The main outcome was pain experience denoted as a Pain Index. A reduction in the Pain Index of 50% from baseline was equivalent to a clinically significant improvement with subsequent classification of the outcome as improved or unimproved. Thirteen of 18 RCTs that met the selection criteria were included in the meta-analysis. The 25 psychological treatments studied in these RCTs included relaxation (11 RCTs), relaxation with biofeedback (four RCTs), cognitive behavioral therapy (nine RCTs), and cognitive behavioral family intervention (one RCT). Twelve RCTs took place in clinic settings and six in school settings. More patients in the treatment group than in the control group had a 50% reduction in the Pain Index from baseline.

A series of articles reported on the treatment of 52 consecutive patients with chronic myofascial pain who had failed to respond to physical, chiropractic, medical, surgical, and pharmacologic treatment with physical therapy combined with EMG biofeedback, counseling, medications, and trigger point injections (Sorrell & Flanagan, 2003; Sorrell, Flanagan, & McCall, 2003). They compared groups with clinically defined anxiety and depression or both with the group having neither. All patients with anxiety took anxiolytic medication during the study, and all but one with depression took antidepressants. Results were that anxiety alone had no effect on outcomes while depressed patients were less likely to improve. Engel, Jensen, and Schwartz (2004) studied three adults with cerebral palsy, using biofeedback-assisted relaxation training on self-reported pain and muscle tension. Two of three participants reported decreases in their pain experiences post-treatment. Their subjective reports, however, did not correspond with physiological changes.

Ninety-two systemic lupus erythematosus (SLE) patients were assigned randomly to receive either biofeedback-assisted cognitive-behavioral treatment (biofeedback/CBT), a symptom-monitoring support (SMS) intervention, or usual medical care (UC) alone (Greco, Rudy, & Manzi, 2004). Biofeedback/CBT participants had significantly greater reductions in pain and psychological dysfunction compared with the SMS group and the UC group. Biofeedback/CBT had significantly greater improvement in perceived physical function compared with UC and improvement relative to SMS was marginally significant. At a nine-month follow-up evaluation, biofeedback/CBT continued to exhibit

relative benefit compared with UC in psychological functioning.

In a study of Complex Regional Pain Syndrome (CRPS), the effects of a multidisciplinary day treatment program were examined by McMenemy, Ralph, Auen, and Nelson (2004). Participants included

11 adults with a history of CRPS of six months or longer. Multidisciplinary treatments used included

physical therapy; occupational therapy; stress management; biofeedback; goal-oriented cognitively based

individual, group, and family counseling; sympathetic blocks; medication management; behavioral

modification; pain management; nutritional education; and case management. Variables assessed at

admission and discharge included physical and occupational therapy ratings, thermal biofeedback levels,

self-reported pain levels, depression and somatic distress levels, narcotic use, and vocation status. At postdischarge

follow up, which ranged from six to 30 months, pain levels, vocational status, and narcotic use were assessed. Results support the hypothesis that multidisciplinary treatment of CRPS is

effective in the

improvement of symptomatology.

Fifty women between 42 and 74 years old with the diagnosis of knee osteoarthritis participated in

a study (Durmus, Alayli, & Canturk, 2005). Patients were randomized into two groups of biofeedback-assisted

isometric exercise or electrical stimulation. For both groups, 20 minutes of therapy was applied five days a week for four weeks. Patients were evaluated before and after therapy. Both

treatment groups

showed significant improvements in pain and physical function scores and demonstrated significant

improvements in anxiety and depression scores.

Phantom limb pain (PLP) was studied in nine individuals (Harden et al. 2005). They received up to seven thermal/autogenic biofeedback sessions over the course of four to six weeks. Interrupted timeseries

analytical models were created for each of the participants, allowing biofeedback sessions to be modeled as discrete interventions. Analyses revealed a 20% pain reduction was seen in five of the nine

patients in the weeks after session four and at least a 30% pain reduction (range: 25 to 66%) was seen in

six of the seven patients in the weeks following session six.

In an illustrative case study, Masters (2006) describes how, after three years of various medical interventions, including exploratory surgery, an individual was referred for biofeedback training.

After a

course of seven sessions over five months that variously included heart rate variability and skin temperature feedback along with extensive home practice of paced breathing and hand warming, the

patient achieved significant symptom reduction and improved coping abilities. A study of 50 chronic pain patients aged 18 to 65 who suffered for at least six months (23 patients with pain in the lumbar region and 27 patients with pain in the cervical and dorsal regions) was reported by Ferrari, Fipaldini, and Birbaumer (2006). The patients were assigned randomly to one of two treatment conditions: 12 sessions of 60 minutes of EMG biofeedback with the electrodes placed in the region of pain and 12 sessions of 80 minutes in a small group. At the end of both treatments, a reduction in the quantity of analgesics consumed, the subjective pain intensity, and the self-evaluations of pain were observed. These improvements continued at the one-month and the six-month follow ups. In a study by Qi and Ng (2007), an eight-week home program provided patellofemoral pain syndrome patients with a treatment with and without EMG biofeedback of the vastus medialis obliquus and vastus lateralis. Twenty-six subjects were randomly allocated into exercise-only or EMGbiofeedback-plus-exercise groups. Both groups performed the same exercise program lasting eight weeks. The intensity of the knee pain was recorded. The results reveal the incorporation of EMG biofeedback into a home exercise program significantly facilitated the activation of the vastus medialis obliquus muscle and the reduction of pain. In a study by Tsai, Chen, Lai, Lee, and Lin (2007), the effects of frontal EMG biofeedbackassisted relaxation on pain in patients with advanced cancer in a palliative care unit was assessed. Participants were randomly assigned to conditions. The experimental group (n = 12) received six EMG biofeedback-assisted relaxation sessions over a four-week period; whereas, the control group (n = 12) received conventional care. The primary efficacy measure was the level of pain, measured by the Brief Pain Inventory. Findings from this study showed frontal EMG biofeedback is effective in reducing cancer-related pain in advanced cancer patients. Voerman, Vollenbroek-Hutten, and Hermens (2006) studied changes in pain, disability, and muscle activation patterns in chronic whiplash (WAD) patients after four weeks of ambulant myofeedback training. Eleven WAD patients received ambulatory myofeedback training, during which upper trapezius muscle activation and relaxation were continuously recorded and processed for four weeks. Feedback was provided when muscle relaxation was insufficient. Pain in neck, shoulders, and

upper back (Visual Analogue Scale), disability (Neck Disability Index), and muscle activation patterns during rest, typing, and stress tasks (surface electromyography) were assessed before and after the four weeks of training. Pain intensity decreased after training. Clinically relevant changes were found with regard to pain in the neck and upper back region and right and left shoulder. A trend for decreased disability was found that was clinically relevant in 36% of the patients. A remarkable reduction was found in the Neck Disability Index items concerning headache and lifting weights.

In a review of studies that evaluated treatments for recurrent abdominal pain (RAP), Weydert, Ball, and Davis (2003) located 10 studies that met the inclusion criteria that the study involve children aged five to 18 years with a diagnosis of RAP, and subjects were randomly assigned to treatment or control groups. Studies that evaluated famotidine, pizotifen, cognitive-behavioral therapy, biofeedback, and peppermint oil enteric-coated capsules showed a decrease in measured pain compared to control groups. The studies that evaluated dietary interventions had conflicting results, in the case of fiber, or showed no efficacy, in the case of lactose avoidance.

In a review of treatment of chronic pain, Singh (2005) reported the therapeutic response of pharmacotherapy in chronic pain at the present time remains unsatisfactory and refractory at best.

Multidisciplinary pain management has not only brought new hope but has also increased the therapeutic response in general. The multidisciplinary management allows patient access to a complete armamentarium of pain therapies and includes relaxation therapy, physiotherapy, transcutaneous electrical nerve stimulation, exercise, biofeedback techniques, acupuncture, behavior modification, hypnosis, sympathetic nerve block, desensitization, and cognition therapy as well as the therapeutic benefit of pharmacotherapy. Multidisciplinary management of chronic pain syndrome has become the key for enhanced success and the route of holistic management.

In a review of mind-body interventions for chronic pain in older adults, Morone and Greco (2007) reported on 20 trials. There was some support for the efficacy of progressive muscle relaxation plus guided imagery for osteoarthritis pain with limited support for meditation and tai chi for improving function or coping in older adults with low back pain or osteoarthritis. In an uncontrolled biofeedback trial that

stratified by age group, both older and younger adults had significant reductions in pain following the intervention. Bohm-Starke, Brodda-Jansen, Linder, and Danielsson (2007) provided 35 women with provoked vestibulodynia four months of treatment with either EMG biofeedback (n=17) or topical lidocaine (n=18).

Assignment to conditions was randomized. Vestibular and general pressure pain thresholds (PPTs) were measured and the health survey Short Form-36 (SF-36) was filled out before treatment and at a six-month follow up. Subjective treatment outcome and bodily pain were analyzed. Thirty healthy women of the same age served as controls for general PPTs and SF-36. Three patients reported total cure, and 25 were improved.

The results of a comprehensive review by the National Institutes of Health Technology Panel are summarized by Lebovits (2007). He reports cognitive-behavioral approaches include hypnosis, relaxation (including guided imagery, progressive muscular relaxation, meditation, and music therapy), biofeedback, coping skills training, cognitive restructuring, supportive and group therapy, and stress-management techniques. The panel concluded the evidence is “strong” (its highest rating) for the effectiveness of relaxation in reducing chronic pain. Specific relaxation strategies that have been shown to reduce levels of pain include guided imagery, progressive muscle relaxation, and meditation. Yet despite the generally accepted efficacy of these methods with pain patients, their relative ease of implementation, and their very low side-effect profile, barriers still exist with the integration of psychological therapies into standard medical care.

In a recent study utilizing EEG biofeedback for Complex Regional Pain Syndrome Type 1 (CRPS-1), Jensen, Grierson, Tracy-Smith, Bacigalupi, and Othmer (2007) reported the results from 18 participants. Pain was measured before and after each 30-minute EEG biofeedback treatment. The EEG biofeedback varied for each participant and across sessions. The authors report a substantial and significant reduction in pain from pre- to post-treatments with 50% reporting clinically meaningful reduction in pain.

In summary, the category of Chronic Pain is a diffuse collection of pain-related, specific disorders, and their treatment with biofeedback techniques has a range of efficacy associated with them.

For many chronic conditions, biofeedback has been shown to be effective in treating pain, especially

when included in a multiple modality program. Therefore, the general conclusion is that biofeedback is efficacious in treating chronic pain, but its utilization for specific disorders needs to be determined from an in-depth review of the literature for that specific condition.

References

- Bohm-Starke, N., Brodda-Jansen, G., Linder, J., & Danielsson, I. (2007). The result of treatment on vestibular and general pain thresholds in women with provoked vestibulodynia. *The Clinical Journal of Pain*, 23(7), 598-604.
- Corrado, P., Gottlieb, H., & Abdelhamid, M.H. (2003). The effect of biofeedback and relaxation training on anxiety and somatic complaints in chronic pain patients. *American Journal of Pain Management*, 13(4), 133-139.
- Durmus, D., Alayli, G., & Canturk, F. (2005). Effects of biofeedback-assisted isometric exercise and electrical stimulation on pain, anxiety, and depression scores in knee osteoarthritis. *Turkiye Fiziksel Tip ve Rehabilitasyon Dergisi*, 51(4), 142-145.
- Engel, J.M., Jensen, M.P., & Schwartz, L. (2004). Outcome of biofeedback-assisted relaxation for pain in adults with cerebral palsy: Preliminary findings. *Applied Psychophysiology and Biofeedback*, 29(2), 135-140.
- Ferrari, R., Fipaldini, E., & Birbaumer, N. (2006). Individual characteristics and results of biofeedback training and operant treatment in patients with chronic pain. *Psicoterapia Cognitiva e Comportamentale*, 12(2), 161-179.
- Flor, H., & Birbaumer, N. (1993). Comparison of the efficacy of electromyographic biofeedback cognitive-behavioral therapy and conservative medical interventions in the treatment of chronic musculoskeletal pain. *Journal of Consulting and Clinical Psychology*, 61(4) 653-658.
- Greco, C.M., Rudy, T.E., & Manzi, S. (2004). Effects of a stress-reduction program on psychological function, pain, and physical function of systemic lupus erythematosus patients: A randomized controlled trial. *Arthritis and Rheumatism*, 51(4), 625-634.
- Harden, R.N., Houle, T.T., Green, S., Remble, T.A., Weinland, S.R., Colio, S., et al. (2005). Biofeedback in the treatment of phantom limb pain: A time-series analysis. *Applied Psychophysiology and Biofeedback*, 30(1), 83-93.
- Hawkins, R.S., & Hart, A.D. (2003). The use of thermal biofeedback in the treatment of pain associated with endometriosis: Preliminary findings. *Applied Psychophysiology and Biofeedback*, 28(4),

279-289.

Humphreys, P.A., & Gevirtz, R. (2000). Treatment of recurrent abdominal pain: Components analysis of

four treatment protocols. *Journal of Pediatric Gastroenterological Nutrition*, 31(1), 47-51.

Jensen, M.P., Grierson, C., Tracy-Smith, V., Bacigalupi, S., & Othmer, S. (2007).

Neurofeedback

treatment for pain associated with complex regional pain syndrome type I. *Journal of*

Neurotherapy,

11(1), 45-53.

Lebovits, A. (2007). Cognitive-behavioral approaches to chronic pain. *Primary Psychiatry*,

14(9), 48-54.

Masters, K.S. (2006). Recurrent abdominal pain, medical intervention, and biofeedback: What happened

to the biopsychosocial model? *Applied Psychophysiology and Biofeedback*, 31(2), 155-165.

McMenamy, C., Ralph, N., Auen, E., & Nelson, L. (2004). Treatment of complex regional pain syndrome

in a multidisciplinary chronic pain program. *American Journal of Pain Management*, 14(2), 56-62.

Morone, N.E., & Greco, C.M. (2007). Mind-body interventions for chronic pain in older adults: A

structured review. *Pain Medicine (Malden, MA)*, 8(4), 359-375.

Newton-John, T.R., Spence, S.H., & Schotte, D. (1995). Cognitive-behavioural therapy versus EMG

biofeedback in the treatment of chronic low back pain. *Behavioural Research & Therapy*, 33(6), 691-697.

Nielson, W.R., & Weir, R. (2001). Biopsychosocial approaches to the treatment of chronic pain. *Clinical*

Journal of Pain, 17(4 Suppl.), S114-S127.

Pulliam, C.B., & Gatchel, R.J. (2003). Biofeedback 2003: Its role in pain management. *Critical Reviews*

in Physical and Rehabilitation Medicine, 15(1), 65-82.

Qi, Z., & Ng, G.Y.F. (2007). EMG analysis of vastus medialis obliquus/vastus lateralis activities in

subjects with patellofemoral pain syndrome before and after a home exercise program. *Journal of*

Physical Therapy Science, 19(2), 131-137.

Singh, A.N. (2005). Multidisciplinary management of chronic pain. *International Medical Journal*, 12(2),

111-116.

Sorrell, M.R., & Flanagan, W. (2003). Treatment of chronic resistant myofascial pain using a multidisciplinary protocol [the myofascial pain program]. *Journal of Musculoskeletal Pain*,

11(1), 5-9.

Sorrell, M.R., Flanagan, W., & McCall, J.L. (2003). The effect of depression and anxiety on the success

of multidisciplinary treatment of chronic resistant myofascial pain. *Journal of Musculoskeletal*

Pain,

11(1), 17-20.

Stinson, J. (2003). Review: Psychological interventions reduce the severity and frequency of chronic pain in children and adolescents. *Evidence-Based Nursing, 6(2), 45-45.* Tsai, P.S., Chen, P.L., Lai, Y.L., Lee, M.B., & Lin, C.C. (2007). Effects of electromyography

biofeedback-assisted relaxation on pain in patients with advanced cancer in a palliative care unit. *Cancer*

Nursing, 30(5), 347-353.

Vlaeyen, J.W., Haazen, I.W., Schuerman, J.A., Kole-Snijders, A.M., & van Eek, H. (1995). Behavioural

rehabilitation of chronic low back pain: Comparison of an operant treatment, an operant-cognitive

treatment, and an operant-respondent treatment. *Clinical Psychology, 34(Pt 1), 95-118.*

Voerman, G.E., Vollenbroek-Hutten, M.M., & Hermens, H.J. (2006). Changes in pain, disability, and

muscle activation patterns in chronic whiplash patients after ambulant myofeedback training. *The Clinical*

Journal of Pain, 22(7), 656-663.

Weydert, J.A., Ball, T.M., & Davis, M.F. (2003). Systematic review of treatments for recurrent abdominal

pain. *Pediatrics, 111(1), e1-11.*

8. Epilepsy

Level 4: Efficacious

Early studies testing EEG biofeedback for epilepsy showed promise in reducing seizure activity, utilizing some form of the technique to increase the abundance of SMR (typically defined at 12-15 Hz)

and often to simultaneously decrease the EEG in the typical low-frequency range of 4-8 Hz. In the first

case study published in 1972, Sterman demonstrated a complete cessation of seizures in a woman who

had a seven-year history of medically uncontrolled generalized tonic-clonic seizures. After becoming

seizure-free, she was issued a state driver's license. This research was an extension of studies with

animals that demonstrated they could be operant-conditioned to increase SMR, and this increase was

associated with an increase in seizure threshold.

Recent studies built on these findings demonstrate self-regulation of slow cortical potentials using

EEG feedback decreases seizure activity in drug-resistant epilepsy when compared to pre-training

(Kotchoubey, Schneider, et al. 1996; Kotchoubey et al. 1999; Sterman, 1986; Swingle, 1998).

This effect

was sustained for at least six months after therapy (Kotchoubey, Blankenhorn, Froscher, Strehl, & Birbaumer, 1997). A five consecutive-day neurobehavioral treatment protocol resulted in 79% of patients being able to achieve seizure control (Joy Andrews, Reiter, Schonfeld, Kastl, & Denning, 2000). Kotchoubey et al. (2001) studied patients with refractory epilepsy in a controlled clinical trial comparing an anticonvulsive drug plus psychosocial counseling (drug), a group that learned to control breathing (control), and a group learning self-regulation of slow cortical potentials (experimental). The experimental and drug groups showed a significant decrease of seizure frequency, but the control group did not.

In a review of the EEG biofeedback treatment for seizures, Sterman (2000) reviewed 18 studies published between 1981 and 1996 in peer-reviewed journals. Most studies used pre-treatment baselines for comparisons, but 10 used appropriate controls such as another biofeedback modality or noncontingent feedback. These trials treated 174 patients with 142 of them (82%) showing clinically significant improvements and 115 of them (66%) demonstrating significant increases in SMR activity. There were no reports of increased seizure activity in those treated with biofeedback. Unfortunately, because none of the studies were designed to be RCTs, this led a Cochrane Database Systematic Review to conclude there is no reliable evidence to support the use of EEG biofeedback in the treatment of epilepsy because of methodological deficiencies and limited number of patients studied (Ramaratnam, Baker, & Goldstein, 2005). However, because most of the subjects were refractory seizure victims, in spite of medication usage, and the biofeedback was shown to clinically reduce the seizure, this technique appears to be effective and safe.

In a recent review by Marson and Ramaratnam (2003), which looked at only RCT studies, one controlled trial was found, and that trial reported significant reductions in median seizure activity. Another review of biofeedback treatment of seizures (Sheth, Stafstrom, & Hsu, 2005) reported a review from 16 studies. Subjects in all studies were designated as having refractory epilepsy. Sample size for most studies was relatively small ($n = 1 - 8$), but one larger sample size study was found ($n = 83$). When all studies were combined, 82% of those treated with biofeedback showed clinical improvement. This

review also presented studies with two other biofeedback techniques, and these are Contingent Negative Variation (CNV) or Slow Cortical Potential (SCP) and Galvanic Skin Response (GSR). Both techniques had positive outcomes with reduction in seizure activity being clinically significant. Pop-Jordanova, Zorcec, and Demerdzieva (2005) report a case study of biofeedback treatment of a 13-year-old girl with psychogenic nonepileptic seizures (PNS). The treatment was electrodermal (EDR) biofeedback combined with cognitive-behavioral therapy. After 10 sessions of 45 minutes per day, they observed cessation of attacks, stabilization of neurotic tendencies, progression of the maturational process, and good academic results. In conclusion, based on more than 30 years of clinical trials with EEG biofeedback based on EEG waveform characteristics for the treatment of seizures, several independent investigators have demonstrated EEG biofeedback is effective in reducing seizure activity, often in refractory patients. There is no evidence this treatment has been linked to an increase in seizures. Other biofeedback techniques (SCP and GSR) have been tried with some success.

References

- Joy Andrews, D., Reiter, J.M., Schonfeld, W., Kastl, A., & Denning, P. (2000). A neurobehavioral treatment for unilateral complex partial seizure disorders: A comparison of right- and left-hemisphere patients. *Seizure*, 9(3) 189-197.
- Kotchoubey, B., Blankenhorn, V., Froscher, W., Strehl, U., & Birbaumer, N. (1997). Stability of cortical self-regulation in epilepsy patients. *Neuroreport*, 27(8), 1867-1870.
- Kotchoubey, B., Schneider, D., Schleichert, H., Strehl, U., Uhlmann, C., Blankenhorn, V., et al. (1996). Self-regulation of slow cortical potentials in epilepsy: A retrial with analysis of influencing factors. *Epilepsy Research*, 25(3), 269-276.
- Kotchoubey, B., Strehl, U., Holzapfel, S., Blankenhorn, V., Froscher, W., & Birbaumer, N. (1999). Negative potential shifts and the prediction of the outcome of neurofeedback therapy in epilepsy. *Clinical Neurophysiology*, 110(4), 683-686.
- Kotchoubey, B., Strehl, U., Uhlmann, C., Holzapfel, S., Konig, M., Froscher, W., et al. (2001). Modification of slow cortical potentials in patients with refractory epilepsy: A controlled outcome study. *Epilepsia*, 42(3), 406-416.
- Marson, A., & Ramaratnam, S. (2003). Epilepsy. *Clinical Evidence*, 9(9), 1403-1420.
- Pop-Jordanova, N., Zorcec, T., & Demerdzieva, A. (2005). Electrodermal biofeedback in treating psychogenic nonepileptic seizures. *Prilozi / Makedonska akademija na naukite i umetnostite*,

Oddelenie

za bioloski i medicinski nauki = Contributions / Macedonian Academy of Sciences and Arts, Section of

Biological and Medical Sciences, 26(2), 43-51.

Sheth, R.D., Stafstrom, C.E., & Hsu, D. (2005). Nonpharmacological treatment options for epilepsy.

Seminars in Pediatric Neurology, 12(2), 106-113.

Sterman, M.B. (1986). Epilepsy and its treatment with EEG feedback therapy. *Annals of Behavioral*

Medicine, 8(1), 21-25.

Sterman, M.B. (2000). Basic concepts and clinical findings in the treatment of seizure disorders with EEG

operant conditioning. *Clinical Electroencephalography, 31(1), 45-55.*

Sterman, M.B., & Friar, L. (1972). Suppression of seizures in an epileptic following sensorimotor EEG

feedback training. *Electroencephalography Clinical Neurophysiology, 33(1), 89-95.*

Swingle, P.G. (1998). Neurofeedback treatment of pseudoseizure disorder. *Biological Psychiatry, 44(11),*

1196-1199.

9. Constipation in Adults

Level 4: Efficacious

A critical review of 38 studies of biofeedback treatment for constipation reported most studies report positive results (Heymen, Jones, Scarlett, & Whitehead, 2003). Success rate for pressure biofeedback (78%) was greater than for EMG biofeedback (70%), but there was no difference in outcome

using intra-anal or perianal EMG sensors. These findings are consistent with another review showing a

62.4% success rate in those treated for constipation (Palsson et al. 2004).

Biofeedback has led to significant improvement in those with constipation (Heymen et al. 1999; Ko et al. 1997; Pucciani et al. 1998). A number of controlled trials have shown EMG biofeedback and

manometry biofeedback had similar effects (Wang, Luo, Qi, & Dong, 2003), biofeedback and electrical

stimulation were comparable (Chang et al. 2003), EMG biofeedback was better than medical treatment

with diazepam or a placebo (Heymen et al. 2007), EMG biofeedback was better than sham biofeedback or

standard care (Rao et al. 2007), and biofeedback was better than laxatives (Chiarioni, Whitehead, Pezza,

Morelli, & Bassotti, 2006). It appears to be more effective for those with pelvic floor dyssynergia than for

those with slow-transit constipation (Bassotti et al. 2004; Battaglia et al. 2004; Chiarioni, Salandini, &

Whitehead, 2005).

Biofeedback has also been used after surgery for rectal disorders. In uncontrolled studies, biofeedback was shown to be of benefit after surgery (Kairaluoma et al. 2004; Hwang et al. 2005; Hwang et al. 2006).

References

- Allgayer, H., Dietrich, C.F., Rohde, W., Koch, G.F., & Tuschhoff, T. (2005). Prospective comparison of short- and long-term effects of pelvic floor exercise/biofeedback training in patients with fecal incontinence after surgery plus irradiation versus surgery alone for colorectal cancer: Clinical, functional and endoscopic/endosonographic findings. *Scandinavian Journal of Gastroenterology*, 40(10), 1168-1175.
- Bassotti, G., Chistolini, F., Sietchiping-Nzepa, F., de Roberto, G., Morelli, A., & Chiarioni, G. (2004). Biofeedback for pelvic floor dysfunction in constipation. *BMJ*, 14, 328(7436), 393-396.
- Battaglia, E., Serra, A.M., Buonafede, G., Dughera, L., Chistolini, F., Morelli, A., et al. (2004). Longterm study on the effects of visual biofeedback and muscle training as a therapeutic modality in pelvic floor dyssynergia and slow-transit constipation. *Diseases of the Colon and Rectum*, 47(1), 90-95.
- Beddy, P., Neary, P., Eguare, E.I., McCollum, R., Crosbie, J., Conlon, K.C., & Keane, F.B. (2004). Electromyographic biofeedback can improve subjective and objective measures of fecal incontinence in the short term. *Journal of Gastrointestinal Surgery*, 8(1), 64-72.
- Brazzelli, M., & Griffiths, P. (2006). Behavioural and cognitive interventions with or without other treatments for the management of faecal incontinence in children. *Cochrane Database of Systematic Reviews (Online)*, 2(2), CD002240.
- Byrne, C.M., Solomon, M.J., Rex, J., Young, J.M., Heggie, D., & Merlino, C. (2005). Telephone vs. face-to-face biofeedback for fecal incontinence: Comparison of two techniques in 239 patients. *Diseases of the Colon and Rectum*, 48(12), 2281-2288.
- Byrne, C.M., Solomon, M.J., Young, J.M., Rex, J., & Merlino, C.L. (2007). Biofeedback for fecal incontinence: Short-term outcomes of 513 consecutive patients and predictors of successful treatment. *Diseases of the Colon and Rectum*, 50(4), 417-427.
- Chang, H.S., Myung, S.J., Yang, S.K., Jung, H.Y., Kim, T.H., Yoon, I.J., et al. (2003). Effect of electrical stimulation in constipated patients with impaired rectal sensation. *International Journal of Colorectal Disease*, 18(5), 433-438.
- Chiarioni, G., Bassotti, G., Stanganini, S., Vantini, I., Whitehead, W.E., & Staganini, S. (2002). Sensory retraining is key to biofeedback therapy for formed stool fecal incontinence. *American Journal of*

Gastroenterology, 97(1), 109-117.

Chiarioni, G., Salandini, L., & Whitehead, W.E. (2005). Biofeedback benefits only patients with outlet

dysfunction, not patients with isolated slow-transit constipation. *Gastroenterology*, 129(1), 86-97.

Chiarioni, G., Whitehead, W.E., Pezza, V., Morelli, A., & Bassotti, G. (2006). Biofeedback is superior to

laxatives for normal transit constipation due to pelvic floor dyssynergia. *Gastroenterology*, 130(3), 657-664.

Croffie, J.M., Ammar, M.S., Pfefferkorn, M.D., Horn, D., Klipsch, A., Fitzgerald, J.F., et al. (2005).

Assessment of the effectiveness of biofeedback in children with dyssynergic defecation and recalcitrant

constipation/encopresis: Does home biofeedback improve long-term outcomes. *Clinical Pediatrics*, 44(1), 63-71.

Davis, K.J., Kumar, D., & Poloniecki, J. (2004). Adjuvant biofeedback following anal sphincter repair: A

randomized study. *Alimentary Pharmacology & Therapeutics*, 20(5), 539-549.

Enck, P., Daublin, G., Lubke, H.J., & Strohmeyer, G. (1994). Long-term efficacy of biofeedback training

for fecal incontinence. *Diseases of the Colon and Rectum*, 37(10), 997-1001.

Fernandez-Fraga, X., Azpiroz, F., Aparici, A., Casaus, M., & Malagelada, J.R. (2003). Predictors of

response to biofeedback treatment in anal incontinence. *Diseases of the Colon and Rectum*, 46(9), 1218-1225.

Fynes, M.M., Marshall, K., Cassidy, M., Behan, M., Walsh, D., O'Connell, P.R., et al. (1999). A prospective, randomized study comparing the effect of augmented biofeedback with sensory biofeedback

alone on fecal incontinence after obstetric trauma. *Diseases of the Colon and Rectum*, 42(6), 753-761.

Guillomot, F., Bouche, B., Gower-Rousseau, C., Chartier, M., Wolschies, E., Lamblin, M.D., et al. (1995).

Biofeedback for the treatment of fecal incontinence: Long-term clinical results. *Diseases of the Colon and Rectum*, 38(4), 393-397.

Heymen, S., Jones, K.R., Ringel, Y., Scarlett, Y., & Whitehead, W.E. (2001). Biofeedback treatment of

fecal incontinence: A critical review. *Diseases of the Colon and Rectum*, 44(5), 728-736.

Heymen, S., Jones, K.R., Scarlett, Y., & Whitehead, W.E. (2003). Biofeedback treatment of constipation:

A critical review. *Diseases of the Colon and Rectum*, 46(9), 1208-1217.

Heymen, S., Scarlett, Y., Jones, K., Ringel, Y., Drossman, D., & Whitehead, W.E. (2007).

Randomized, controlled trial shows biofeedback to be superior to alternative treatments for patients with pelvic floor dyssynergia-type constipation. *Diseases of the Colon and Rectum*, 50(4), 428-441.

Heymen, S., Wexner, S.D., Vickers, D., Nogueras, J.J., Weiss, E.G., & Pikarsky, A.J. (1999). Prospective, randomized trial comparing four biofeedback techniques for patients with constipation. *Diseases of the Colon and Rectum*, 42(11), 1388-1393.

Hibi, M., Iwai, N., Kimura, O., Sasaki, Y., & Tsuda, T. (2003). Results of biofeedback therapy for fecal incontinence in children with encopresis and following surgery for anorectal malformations. *Diseases of the Colon and Rectum*, 46(10 Suppl), S54-8.

Hwang, Y.H., Choi, J.S., Nam, Y.S., Salum, M.R., Weiss, E.G., Nogueras, J.J., et al. (2005). Biofeedback therapy after perineal rectosigmoidectomy or J pouch procedure. *Surgical Innovation*, 12(2), 135-138. 23(2), 189-194.

Ryn, A.K., Morren, G.L., Hallbook, O., & Sjordahl, R. (2000). Long-term results of electromyographic biofeedback training for fecal incontinence. *Diseases of the Colon and Rectum*, 43(9), 1262-1266.

Shafik, A., El Sibai, O., Shafik, I.A., & Shafik, A.A. (2007). Stress, urge, and mixed types of partial fecal incontinence: Pathogenesis, clinical presentation, and treatment. *The American Surgeon*, 73(1), 6-9.

Solomon, M.J., Pager, C.K., Rex, J., Roberts, R., & Manning, J. (2003). Randomized, controlled trial of biofeedback with anal manometry, transanal ultrasound, or pelvic floor retraining with digital guidance alone in the treatment of mild to moderate fecal incontinence. *Diseases of the Colon and Rectum*, 46(6), 703-710.

Sunic-Omejc, M., Mihanovic, M., Bilic, A., Jurcic, D., Restek-Petrovic, B., Maric, N., Dujsin, M., & Bilic, A. (2002). Efficiency of biofeedback therapy for chronic constipation in children. *Collegium Antropologicum*, 6 (Suppl), 93-101.

Terra, M.P., Dobben, A.C., Berghmans, B., Deutekom, M., Baeten, C.G., Janssen, L.W., Boeckxstaens, G.E., Engel, A.F., Felt-Bersma, R.J., Slors, J.F., Gerhards, M.F., Bijnen, A.B., Everhardt, E., Schouten, W.R., Bossuyt, P.M., & Stoker, J. (2006). Electrical stimulation and pelvic floor muscle training with biofeedback in patients with fecal incontinence: A cohort study of 281 patients. *Dis Colon Rectum*, 49(8), 1149-1159.

van Ginkel, R., Benninga, M.A., Blommaart, P.J., van der Plas, R.N., Boeckxstaens, G.E., Buller, H.A., et al. (2000). Lack of benefit of laxatives as adjunctive therapy for functional nonretentive fecal soiling in children. *Journal of Pediatrics*, 137(6), 808-813.

Wang, J., Luo, M.H., Qi, Q.H., & Dong, Z.L. (2003). Prospective study of biofeedback retraining in patients with chronic idiopathic functional constipation. *World Journal of Gastroenterology*, 9(9), 2109-2113.

Wiesel, P.H., Norton, C., Roy, A.J., Storrie, J.B., Bowers, J., Kamm, M.A. (2000). Gut focused behavioural treatment (biofeedback) for constipation and faecal incontinence in multiple sclerosis. *Journal of Neurology, Neurosurgery, and Psychiatry*, 69(2), 240-243.

SECTION C

Cost Effectiveness

The above sections demonstrate that many biofeedback based interventions are highly efficacious and that some have been shown to work as well or better than drugs. In addition, they don't have any side effects.

A common assumption is that behavioral interventions must cost far more than taking medications as ten or more sessions of biofeedback (perhaps as many as 40 for some EEG based interventions) are needed to achieve lasting effectiveness and each session costs a minimum of \$150 (frequently \$300 or more) in private pay situations. Biofeedback based interventions have been shown to help only an average of 80% of patients at all with the average amount of help also being about 80% control of symptoms. Thus, at least 20% of patients will need to get other types of therapy and the third party payer or the patient is out the cost of the biofeedback therapy.

However, medications for the conditions discussed above seldom help any more patients any better than biofeedback. So, the same situation holds for medications as for biofeedback.

In the mid 1980s experts such as Carol Schneider* began trying to figure out the relative cost of biofeedback vs. medications when all the factors were taken into account.

[*Biofeedback Self Regul.](#) 1987 Jun;12(2):71-92. Cost effectiveness of biofeedback and behavioral medicine treatments: a review of the literature. [Schneider CJ.](#)

Carol and many others since then have come up with the same basic conclusion: Biofeedback – even when augmented with relaxation training and / or cognitive restructuring actually costs less in the long run than medicine based treatments.

Here are some of the key factors:

1. Medications can be incredibly expensive and insurance frequently doesn't cover much or all of the cost of sufficient numbers of pills.
2. It is very rare that the correct medication is picked the first time a patient sees a primary care provider and doses need to be tinkered with ad infinitum.
3. Many – if not most – patients wind up making multiple visits to incredibly expensive sub-specialists such as neurologists.

4. The visits never stop because the drugs usually have to be changed and doses titrated as the drugs lose effectiveness over time.
5. Side effects are frequently expensive to treat.
6. Sometimes drug effects require hospitalization and rehospitalization.
7. Drugs have a poor record of long term relief from the disability related to chronic pain. This disability (e.g. days of work lost) costs the patient and society a fortune.
8. Reviews such as Schneider's consistently indicate that "that multicomponent behavioral medicine treatments are cost-effective on all dimensions reviewed.
9. Cost/benefit ratios range between 1:2 and 1:5, with a median of 1:4."
10. Often, there may not be any really effective drugs for some conditions such as IBS.
11. Many patients do not respond to any medications for their conditions or respond minimally.
12. Patients frequently won't take potentially effective medications due to significant side effects.

Herman PM, Craig BM, Caspi O. BMC Complement Altern Med. 2005 Jun 2;5:11.
Is complementary and alternative medicine (CAM) cost-effective? A systematic review.

Insurance coverage:

For current information on insurance coverage for biofeedback, see Ron Rosenthal's article "New Guidelines for Third Party Reimbursement for Biofeedback" at www.aapb.org.

Summary of reasons to refer patients for bfb

1. Biofeedback is not magic. Rather, muscle tension biofeedback helps patients learn to recognize incorrect patterns of muscle tension proven to cause headaches and to correct those patterns. Same idea for temperature.
2. Biofeedback is not "experimental" when used for headache control. Rather, biofeedback for prevention of tension headaches and migraine headaches of non-traumatic origin has about as much solid research supporting its efficacy as that available for most preventive headache medications.
3. Biofeedback has been shown to have effect sizes proving it to be as efficacious as popular preventive medications.

It helps at least the same percent of patients to about the same extent as medications (about 80% of patients get 80% better).

4. Biofeedback has no side effects.
5. It lasts for up to 15 years.
6. It is cost effective.

Modified from Andrasik's slides 2013 – used with his permission.

For those practitioners who prefer to prescribe medications rather than send patients for behavioral interventions:

When to choose behavioral treatments over medications

- Patient prefers a non-drug approach
- Drug treatment cannot be tolerated or is medically contraindicated
- Response to drug treatment is absent or minimal

- Patient is pregnant, has plans to become pregnant, or is nursing
- History of frequent or excessive use of analgesic or other acute medications
- Significant life stress or deficient stress-coping skills

Section L

Further reading and References

A. Further reading

This book is oriented toward helping clinicians learn to perform relatively simple, clinically oriented studies. As such, it certainly doesn't have the depth in specialized design strategies, outcome measures, or statistics to provide the information needed for many types of advanced studies. When you find that you are working beyond the depth of this book, you probably need to seek help from research professionals and clinicians highly experienced in performing research in the field you want to tackle. Before talking to them, looking over several of the following texts may be helpful so you have an idea of the usual approaches attempted for your type of project. They will be even more helpful after your conversations because you can look up the mass of unfamiliar design and test names that probably swamped you during the discourse.

1. Design / Methodology:

a. Bordens K and Abbott B: *Research design and methods*; Second edition, Mayfield, London 1991.

b. Hulley S, Cummings S: *Designing Clinical Research: An epidemiologic approach*. Williams & Wilkins, Baltimore; 1988.

c. Troidl H: *Principles and Practice of Research: Strategies for Surgical Research*. Springer-Verlag, New York, 1991.

2. Data and Statistics:

- a. Spilker B: Guide to clinical interpretation of data. Raven Press, New York, 1986.
- b. McCall R: Fundamental statistics for behavioral sciences. Harcourt Brace of Philadelphia, 1994.
- c. Hays W: Statistics. Holt, Rinehart and Winston of New York, 1988

3. Outcome measures:

- a. The American Academy of Orthopaedic Surgeons: Fundamentals of Outcome Research, American Academy of Orthopaedic Surgeons; (708) 823-7186; 6300 North River Road; Rosemont, Illinois 60018.
- b. Pynsent P, Fairbank J, Carr A: Outcome measures in orthopaedics. Butterworth-Heinemann, Oxford, 1993.

4. Writing and Presentations:

- a. James D: Writing and speaking for excellence. Jones and Bartlett publishers of Boston, 1996.
- b. Bordens K and Abbott B: Research design and methods; Second edition, Mayfield, London 1991.

B. References

Belmont Report: Ethical principles and guidelines for the protection of human subjects of research. OPRR Reports, US Government Printing Office 2010778-80319, 1988.

Beyerstein B: Why bogus therapies seem to work. Skeptical Inquirer 21: 29 - 34, 1997.

Bordens K, Abbott B: Research Design and Methods: A Process Approach, Second Edition. Mayfield Publishing Company, Mountain View California, 1991.

Broad W, Wade N: Betrayers of the truth: Fraud and deceit in the halls of science. Simon and Schuster, New York, 1982.

Brownlee, S: Bad science + breast cancer. Discover 73 - 78, August 2002.

Clark E: A preliminary investigation of the neoprene tube finger extension splint. Journal of Hand Therapy 10: 213-221, 1997.

Condon E: Scientific study of unidentified flying objects. Vision, 1968

Crano W, Brewer M: Principles and methods of social research, Allyn and Bacon, Boston, 1986

Dingell J: Misconduct in Medical Research. New England Journal of Medicine 238: 1610-1615, 1993.

Einhorn T, Burstein A, Cowell H: Human Experimentation. JBJS 79A: 959 - 960, 1997.

Egner, T, Strawson E, Gruzelier J: EEG Signature and phenomenology of alpha/theta neurofeedback training versus mock feedback. Applied Psychophysiology and Biofeedback 27: 261 - 270, 2002.

Fuson R, Sherman M, Van Vleet J, Wendt T: The conduct of Orthopedic clinical trials. JBJS 79A: 1089 - 1098, 1997.

Garland J: JBJS 70: 1357, 1988.

Greenhalgh T: How to read a paper: The basics of evidence based medicine. BMJ Publishers, London, 1997.

Hammond C: Science, Pseudoscience, and Politics. Journal of Neurotherapy 6: 1-6, 2002

Heuvelmans B: In the wake of the sea-serpents. Hill and Wang, New York, 1968

- Hulley S, Cummings S: Designing Clinical Research: An epidemiologic approach. Williams & Wilkins, Baltimore; 1988
- Kahn C: Picking a research problem. New England Journal of Medicine, 330:1530-1533, 1994.
- Kane, M: Research Made Easy in Complementary and Alternative Medicine. London, Churchill Livingstone, 2004.
- Kerluke L, McCabe S: Journal of Hand Surgery (Amer.) 18: 1-3, 1993.
- Knight J: Exploring the compromise of ethical principles in science. Perspectives in Biology and Medicine 27: 432 - 442, 1984.
- Levine R: Ethics and regulation of clinical research. Urban & Schwarzenberg, Baltimore, 1986
- McQuay H, Carroll D, Moore A: Pain 64: 331 – 335, 1995.
- Monagle J, Thomasma D: Health care ethics. Aspen, Maryland, 1994.
- Moser C, Kalton G: Survey methods in social investigation. Basic Books, New York, 1972.
- Ogden T: Research Proposals: A guide to success. Raven Press, N.Y. 1991.
- Ostle B: Appendix 10 to Statistics in Research, Iowa State University Press, Ames, Iowa, 1963.
- Pedhazur E, Pedhazur-Schmelkin L: Measurement, Design, and Analysis. New Jersey, Lawrence Elbaum Associates, 1991.
- Pollard R, Katkin E: Placebo effects in biofeedback and self-perception of muscle tension. Psychophysiology 21: 47 – 53, 1984.
- President's Commission for the study of ethical problems in medicine and biomedical and behavioral research: Whistle-blowing in biomedical research: Policies and procedures for responding to reports of misconduct. US Government Printing Office stock # 040-000-00471-8, Washington, 1982.
- Price D, Milling L, Kirsch I, Duff A, Montgomery G, Nicholls S: An analysis of factors that contribute to the magnitude of placebo analgesia in an experimental paradigm. Pain 83: 147 – 156, 1999.
- Pynsent P, Fairbank J, Carr A: Outcome measures in orthopedics. Butterworth-Heinemann, Oxford, 1993.
- Rousseau D: Case studies in pathological science. American Scientist 80:54-63, 1992.

Sherman R: Potential regulation of biofeedback devices and practice by the FDA. Biofeedback 22(3): 6 - 8, 1994

Sherman R: Pain Assessment and Intervention from a Psychophysiological Perspective. Association for Applied Psychophysiology and Biofeedback, Colorado Springs, 2004.

Sherman R, Arena J: Biofeedback in the assessment and treatment of low back pain. Chapter 8 in: (J.V. Basmajian and R. Nyberg, eds) Rational Manual Therapies. Williams & Wilkins, pages 177 - 197, 1992.

Sherman R, Camfield M, Arena J: The effect of presence or absence of pain on low back pain patients' answers to questions on the MMPI's Hy, Hs, and D scales. Journal of Military Psychology, 7(1): 28-38, 1995.

Sherman R, Devor M, Jones C, Katz J, Marbach J: Phantom pain, New York, Plenum Press; 1996.

Sherman R, Heath G: Changes in finger tip temperature during extended baselines and across days. Biofeedback, 26:28-31,1998.

Sherman R, Sherman C, Gall N: A survey of current phantom limb pain treatment in the United States. Pain 8: 85 - 99, 1980.

Silverman F: Research design in speech pathology and audiology. Prentice-Hall, New Jersey, 1977.

Smith L, Oldman A, McQuay H, Moore R: Teasing apart quality and validity in systematic reviews: An example from acupuncture trials in chronic neck and back pain. Pain 86: 119 – 132, 2000.

Spence R, Shimm D, Buchanan A: Conflicts of interest in clinical practice and research. Oxford University Press, New York, 1996.

Spilker B: Guide to clinical interpretation of data. Raven Press, New York, 1986.

Troidl H: Principles and Practice of Research: Strategies for Surgical Research Springer-Verlag, New York, 1991.

Turk D, Melzack R: Handbook of pain assessment. Guilford press, New York, 1993.

Vertosick F: First, do not harm. Discover, 106 - 110, July 1998:

Publishing and Copying Information

Clinical Research: Skills clinicians need to maintain effective practices by Richard A. Sherman, PhD is published by The Behavioral Medicine Research and Training Foundation, 6576 Blue Mountain Road, Port Angeles, WA 98362. Copyright 2007 by The Behavioral Medicine Research and Training Foundation. All rights reserved. Do not copy or otherwise reproduce any portion of this book without written permission from the author.

About the Author

Rich Sherman received his doctorate in psychobiology over a quarter of a century ago. For most of that time, he has taught at the undergraduate and graduate levels and mentored all levels of students from high-schoolers getting their first peek at clinical research through medical fellows and staff doing complex clinical projects. He regularly teaches classes and seminars in research methodology, statistics, and pain mechanisms and treatment. He has been director of clinical psychophysiology laboratories at three medical centers and has concurrently been chief of orthopedic and surgical research at two of them. These years of watching residents and other students cut their teeth on research have given him a pragmatic approach to getting across the crucial information in ways students can relate to.

