

# Statistics and Clinical Trials: Past, Present and Future

Herbert I. Weisberg  
Correlation Research, Inc.  
61 Pheasant Landing Road, Needham, MA 02492

## Abstract

The rationale for large-scale randomized clinical trials so predominates today that essential limitations of this gold standard are rarely considered. However, this attitude is quite recent and in sharp contrast with earlier notions about the nature of clinical research. Acceptance of human experimentation in its current form came about largely because of particular conditions that pertained shortly after World War II. Although current circumstances have changed, the basic methodology of clinical trials has remained essentially constant. However, new realities imply the need to reconsider the ways in which these studies should be designed and analyzed in the future.

**Key Words:** Clinical trials, history of medicine, potential outcomes, individualized treatment effects, personalized medicine

## 1. Introduction

The recent blockbuster movie *Rise of the Planet of the Apes* closes ominously as a drop of blood drips from the brow of an airline pilot who has contracted a deadly virus. This virus is the delivery vehicle for a wonder drug intended to cure Alzheimer's. It works wonders for simian intelligence, but not so well in humans, resulting ultimately in the demise of humanity. Few viewers I have surveyed seem bothered by an inconsistency with the 1968 original *Planet of the Apes*, in which a nuclear holocaust was responsible for our downfall and the apes' subsequent ascendancy. It seems as if biological agents, especially those developed by evil and greedy drug companies, have increasingly displaced nuclear weapons as the Frankenstein monsters in our cinematic nightmares.

While movies are now in living color, the effects of drugs are portrayed as completely black and white. The drug is either a hero, capable of miracle cures for all who are treated, or an arch-villain, wantonly destroying its unwitting victims. Unfortunately, these caricatures are not all that far from the way scientists and regulators tend to regard pharmaceutical products. Is it time to adopt a more nuanced view of how drugs and medical devices work and how they should be evaluated?

## 2. A Brief History

For most of human history, medical practice was an art practiced by "healers" and based on esoteric knowledge acquired primarily through apprenticeship. The value of this healing was believed to inhere primarily in the relationship between physician (or shaman, herbalist, etc.) and patient rather than in any specific standardized treatment modality. The actual benefits and harms resulting were unclear and probably highly

variable over time, geographic areas, societies, etc. Presumably, the experience and skill of the individual practitioner was mainly responsible for any real or perceived efficacy of the “treatments” rendered.

During the 1700’s, in the Age of Enlightenment, things gradually began to change. The scientific method, based on empirical evidence about diseases and the impacts of different interventions, began to be applied. However, with the major exception of disputes over the wisdom of inoculation to prevent smallpox, the study of medical treatment remained almost entirely qualitative. At the end of the eighteenth century, Gilbert Blane described the research process as hinging almost exclusively on clinical reasoning, with no hint of quantitative analysis:

There is . . . a great difficulty attending all practical inquiries in medicine; for in order to ascertain truth, in a manner that is satisfactory to a mind habituated to chaste investigation, there must be a series of patient and attentive observations upon a great number of cases, and the different trials must be varied, weighed, and compared, in order to form a proper estimate of the real efficacy of different remedies and modes of treatment (Blane, 1785).

Blane would not have understood the relevance of statistical data to treatment efficacy. Not too long after, however, statistical ideas would be in the air. By the early nineteenth century, the concept of probability, essentially in its current form, was well known. Pierre-Simon de Laplace was the leading scientific theorist of the era, and a strong believer in the potential of statistical analysis in various fields. In medical research, he advocated comparing rates of success between alternative therapeutic interventions:

The probability calculus can make one appreciate the advantages and disadvantages of the methods used in the speculative sciences. Thus, to discover the best treatment to use in curing a disease, it is sufficient to test each treatment on the same number of patients, while keeping all circumstances perfectly similar. The superiority of the most beneficial treatment will become more and more evident as this number is increased, and the calculus will yield the corresponding probability of its benefit and of the ratio by which it is greater than the others (Laplace, 1825).

Laplace’s prescription was mainly theoretical, but influenced some contemporary medical researchers, who saw statistical comparisons as a potential improvement on qualitative clinical evaluations. One of the leaders in promoting quantitative analysis was Pierre Louis, who formulated the “numerical method” of assessing treatment efficacy. His approach, applied first in the 1820’s, was primitive by modern standards, utilizing simple counts without formal probabilistic analysis. However, his results inspired others, primarily in France, to begin seeing the potential of statistical methods.

The ideas implemented by Louis and some others to count, and even sometimes to compare, rates of successful outcomes for different therapies seems an obvious step to us, but was in fact quite controversial. The main sticking point had to do with the conception of medical practice that was prevalent at the time. Most physicians, even research-minded ones, regarded their practice to be as much art as science. They viewed statistics as

glossing over the multitude of specific factors that characterized the individual patient and to which their “medical tact” would be applied. Counting outcomes was in their eyes like treating an “average man,” a mythical being made famous by Adolphe Quetelet. Moreover, the numbers of individuals necessary for valid statistical conclusions was essentially unknown and thought to be extremely large (remember this was well before modern notions of significance and confidence intervals).

Despite these limitations, advocates of the numerical method had some strong arguments too. They pointed out that homogeneity of patients might be good enough to obtain reasonable conclusions. Even if imperfect, empirical data could go far toward debunking the overconfidence many doctors felt about the conclusions they could derive from their personal experience. Some objective evidence was necessary, they argued, to place medicine on a more scientific footing. In effect, the controversy revolved around the relative trustworthiness of two ways to make medicine more scientific: rely primarily on the detailed examination of individual circumstances by expert practitioners or on statistical evidence from collections of many superficially similar cases. At that time, it was a choice between two highly imperfect methodologies.

The issue of how to regard numerical evidence came to a head in Paris in the mid-1830's. A statistical study comparing an innovative surgery technique for removal of bladder stones with a traditional one had been performed by the distinguished surgeon Jean Civiale. His data appeared to demonstrate a dramatic advantage in survival for the new approach, lithotripsy. However, the medical establishment remained divided. Civiale's landmark study precipitated a debate that raged for several years about the role of the numerical method in the scientific evaluation of therapies.

Pierre Louis and Simeon-Denis Poisson carried the flag for probability and statistics, although Poisson was surprisingly cautious at the outset. Poisson, of course, was the leading disciple of Laplace and was in the process of producing a classic work on applications of probability to judgments in civil and criminal matters. The prominent physicians Francois Double and Risueno d'Amador led the opposition. The battle was joined, but no clear resolution emerged. Pragmatists at the time recognized that statistical evidence was important to consider, but were unsure just how much weight the numerical data ought to be given.

In the aftermath of this imbroglio, a few pioneers, such as Jules Gavarret in France and his student Elisha Bartlett in America, continued to push for quantification. They appreciated the limitations of statistical comparisons, but advocated their use when cases were deemed sufficiently similar and large numbers could be obtained. However, in general, the impetus for the numerical method seemed to peter out. Why were the (to us) obvious advantages of statistical methods not recognized until much later? There were at least three important explanatory factors.

One factor was simply the primitive state of data collection and analysis, which left many critics doubtful. For example, in the debates over lithotripsy, the quality (reliability, objectivity, consistency, etc.) and quantity (number of observations) of the data were questioned. As for the sample sizes required, the standard of proof suggested by Poisson was stringent, essentially a significance level of .005. Few studies could satisfy this demanding criterion. Even strong supporters of the numerical method were careful to explain that the conditions necessary for calculating averages and comparing them to obtain evidence of relative efficacy were not easily satisfied.

Part of the difficulty for statisticians of the day was the entanglement of two questions that have been separated, and presumably resolved by modern theory. First, how large must the sample sizes be in order to be confident that the observed average difference is real? Second, how can the multitude of possible underlying causes (confounding factors) be taken into account? William Guy, a leading British applied statistician and practitioner of forensic medicine, was guardedly supportive of the numerical method. His reservations stemmed from the belief that very large numbers of “facts” (i.e. observations) would be required for scientific proof and that the calculus of probability could not be the sole determining factor, even theoretically, of the necessary sample sizes.

I now proceed to inquire what number of facts in any given inquiry will suffice to yield a true average; and what use are we allowed to make of averages derived from comparatively small numbers of facts. It is not easy to answer this question. It is probable, and even certain, that the number will vary with the nature of the facts themselves, and that, as a general rule, more facts will be necessary in the case of units of variable magnitude than in the case of simple units; more facts in the case of events brought about by the combined action of many forces than...by a small number of forces acting together; more facts when, as in such a question as the duration of human life, we admit into our tables several classes than when we restrict our inquiry to one class.  
(Guy, 1860)

Guy was well aware that probability theory could be applied to the problem of determining sample sizes. He explicitly rejected this approach.

This question is from the very nature of the case insusceptible of mathematical treatment; but you are probably aware that there are mathematical formulae for calculating the limits of possible error attaching to any given number of facts, large or small, irrespective of the nature of the facts. Of these mathematical formulae Gavarret has made much use in questioning the sufficiency in point of number of some collections of facts made by Louis. ...But such applications of the pure mathematics must be very rare; and are certainly not free from objections.  
(Guy, 1860)

A second factor was the general disillusionment with probability theory as a tool for reasoning about human actions and behavior. Poisson and his followers, elaborating on earlier attempts by Laplace and Condorcet, were applying the theory of probability to voting and judicial decision-making. By around 1840, these efforts came to be generally seen as over-reaching, resulting in a severe backlash. The philosopher John Stuart Mill in 1843 famously characterized these applications of probability theory as “the real opprobrium of mathematics.” In this climate, skepticism on the part of the medical profession about the value of statistics was undoubtedly reinforced.

Finally, opposition to statistical methods came from another quarter that might seem to us very surprising. Statistical methods were seen by some as being antithetical to the scientific method. Most outspoken and influential was Claude Bernard, the father of modern physiology. Bernard believed that medical science should be strictly deterministic, and that progress could result only from a deeper understanding of how the various organs and biological processes functioned. He had no use for statistics and taught that each individual was a unique organism. Grouping together such diverse individuals in an average was not helpful to the physician. Indeed, he felt it was unscientific and regressive, a throwback to the pure empiricism that must ultimately give way to deterministic knowledge of general laws derived from anatomy, physiology and biochemistry.

The results of statistics, even statistics of large numbers, seem indeed to show that some compensation in the variations of phenomena leads to a law; but as this compensation is indefinite, even the mathematicians confess that it can never teach us anything about a particular case.  
(Bernard, 1865)

Ironically, Bernard himself was helping to lay the groundwork for a future revival of statistical methods in medicine. In 1862, Bernard had collaborated with Louis Pasteur on the first study of a new process for preserving wine that later became known as pasteurization. This work led in the 1860's to the germ theory of disease. The discovery that single causes (micro-organisms) could be isolated was a powerful vindication of lab-based biological research. However, this cause operated in a way that was essentially independent of individual characteristic and circumstances. So, the effect of a measure to eliminate this cause was amenable to relatively straightforward statistical analysis.

For example, as early as 1870, Joseph Lister published a landmark study on the effectiveness of antiseptic surgery. He compared the mortality rate in the University of Edinburgh Hospital during 1867-1869 (after antiseptic methods) with the historical experience during 1864-1866 and showed a sharp decline. However, the prevailing ambivalence about statistics induced Lister to emphasize the importance of the (deterministic) theory of Pasteur as the major element in the scientific proof supporting sterilization.

This pragmatic but reserved attitude toward statistical evidence continued to hold sway for many years. Even the major advances in statistical thinking initiated by Francis Galton and Karl Pearson between 1890 and 1920 had little impact on clinical research. Most medical doctors failed to understand or perceive the relevance of statistics, other than as (sometimes) a confirmation of what they already knew. The application of mathematical probability theory would have seemed to most of them ludicrously abstruse. For them, medical science could advance by careful recording of clinical case studies and occasional breakthroughs in fundamental biological research. Furthermore, the statistical methods being developed by Pearson, Edgeworth, and Yule were primarily concerned with large samples that were impractical in much medical experimentation.

### 3. Emergence of the Randomized Clinical Trial

Prior to 1920, statistical comparisons of medical interventions were primarily observational and based on data collected retrospectively, such as that analyzed by Lister in his study of antiseptic surgery. A few pioneering prospective studies were undertaken, but the control groups, if any, were largely dictated by expediency. For example, an important study of an anti-typhoid vaccine developed by Almroth Wright in 1896 was based on a comparison of volunteers from the British Army with other soldiers in the same regiments. Efforts were sometimes made to find broadly “similar” groups for these comparisons, but questions about comparability and potential bias always remained. Moreover, the ethical dilemma posed by the need to withhold a potentially life-saving treatment from the control subjects was disquieting, as dramatically portrayed in the novel *Arrowsmith*, written by Sinclair Lewis in 1925.

Two major scientific advances that both took place in the 1920's laid the groundwork for a sea change in this situation. In medicine, the fortuitous discovery of penicillin in 1928 led to a proliferation of new antibiotics that transformed medical practice. Administration of these life-saving treatments was relatively straightforward and depended little if at all on the subtle medical judgment on which physicians prided themselves. The drugs' effects were also much less variable with respect to the patient's response than most traditional therapies. These simplifications reduced the force of arguments that emphasized the diversity of individuals and questioned the interpretability of simple percentages. However, knowledge of statistical methods among medical professionals grew slowly. Clinicians were trained to deal with individuals, not populations, and placed great faith in their professional judgment, as did the general public.

In statistics, R. A. Fisher published *Statistical Methods for Research Workers* in 1925. Fisher's revolutionary innovations in experimental design and analysis proved enormously successful in several fields, especially in agricultural research and industrial engineering. Fisher's methods of significance testing, building upon W.S. Gossett's initial breakthrough of the t-test in 1908, finally solved the problem of how to analyze experiments with modest sample sizes. His designs based on random assignment of treatments to experimental units provided not only a firm basis for calculating p-values for significance tests, but also a means of eliminating bias resulting from uncontrolled “causes” influencing the outcome. At one brilliant stroke, Fisher had untangled and resolved these two major difficulties that had hamstrung statistical comparisons throughout the nineteenth century.

Fisher's ideas were widely adopted and further developed by many other researchers and mathematicians. One very important but controversial direction of this development was initiated by Jerzy Neyman and E.S. Pearson. These pioneers regarded statistics primarily as a way to guide decisions, an approach that Fisher found inappropriate for scientific research. However, the Neyman-Pearson decision-theoretic methods, including such concepts as confidence intervals, Type I and Type II error rates and statistical power, became widely accepted. These tools proved particularly useful for research related to industrial product development, production and testing. However, the use of statistics to evaluate the efficacy and safety of medical interventions remained quite limited.

It was not until after World War II that conditions were ripe for the application of statistical methods, and randomized experimentation in particular, to evaluate medical treatments. During wartime, an extraordinary level of cooperation among researchers had been orchestrated by the U.S. and British governments in order to address pressing

problems, especially those pertaining to battlefield casualties and illnesses. To some extent, the organizational structures created to meet this need carried over into the post-war years. At the same time, Austin Bradford Hill began promoting the use of randomized experimentation, which had been so successful in other domains. While the technical problems of dealing with modest samples and controlling for confounding had been basically solved, serious practical problems and psychological resistance had yet to be overcome. Most practicing physicians balked at the idea of random assignment and “blinding.”

The first randomized clinical trial, directed by Hill in 1946 under the auspices of the British Medical Council, demonstrated clearly the benefits conferred by streptomycin for the treatment of tuberculosis. Hill’s pioneering efforts began shifting medical opinion toward a greater appreciation of the value to society of definitive findings that could be obtained via RCTs. Obtaining knowledge for future application must be considered part of the physician’s responsibility, along with the primary task of curing each individual. Within a few years, many researchers were following in Hill’s footsteps and the golden age of the RCT had begun.

#### **4. Clinical Trials Today**

By the 1970’s, the methodology of the large-scale, double-blinded, randomized controlled trial had essentially reached maturity. This approach was broadly accepted as the only way to “prove” the efficacy of a pharmaceutical product or medical device, and was mandated by regulatory agencies throughout the world. The RCT has since become the fulcrum around which revolves a mammoth pharmaceutical industry. Clinical research organizations (CROs) and software development companies have evolved to meet the specialized organizational and technical requirements of these studies. Literally billions of dollars can be involved in the complex and lengthy process of developing and obtaining regulatory approval, with the Phase III RCT results as the critical determining factor.

There is no question that the modern RCT has made an enormous contribution to human health and wellbeing. Replacing clinical judgment with “proof” based on randomization and statistical significance has been a major step forward. A plethora of new drugs and devices have generated huge profits for their developers and suppliers, while alleviating suffering and reducing mortality for their users. It is not surprising that commercial enterprises built around the methodological engine driving this cornucopia of benefits should continue to expand and thrive. It is virtually inconceivable today to imagine how things could possibly be different. Yet, there are critics who voice an array of troubling concerns.

Increasingly, the huge investments of time and expense required by large-scale RCTs fail to deliver the definitive answers anticipated. Promising results from one study often cannot be replicated in subsequent research. Products that appear safe and effective in trials leading to approval may turn out to have problems that are only discovered after much wider application post-approval. Some treatments that seem firmly established are later called into question, either for safety concerns or an apparent loss of efficacy. These and other sources of ambiguity can make it difficult for the medical practitioner to interpret and apply a study’s findings to individual patients. Indeed, there is often a

disconnect between the black-and-white “decisions” offered by the statistical paradigm and the shades of gray that confront clinicians in dealing with individual patients.

At a more global level, some skeptics are concerned about the “medicalization” of society. They question the development and aggressive marketing of products for which lifestyle changes, or perhaps some form of alternative and complementary medicine, might be effective. Many people who would not have been considered ill a generation ago are consuming multiple powerful medications on a regular basis. The benefits they receive should ideally be weighed against potential adverse effects and drug interactions that are often not well catalogued. Again, the RCTs provide little guidance to help the practicing physician sort through these complexities.

## 5. What is the Real Problem?

Some critics of the RCT have pointed mainly to practical difficulties in implementing its stringent requirements: randomization, blinding of investigators and participants, long-term follow-up, subject retention, accurate data collection, etc. The usual response is that these technical problems can be overcome by scrupulous (and extremely costly) adherence to proper research practice. Criticisms of RCTs fall on deaf ears because they are tilting against a platonic ideal of research methodology. *Theoretically*, the RCT appears to be unimpeachable; its internal logic is compelling and promises to deliver unequivocal results. Yes, there are practical problems argue its defenders, but nothing is perfect. Look at the great successes that have been achieved for decades using the RCT. Besides, what else can we do?

Regrettably, discussions about methodology are hampered by an overly narrow view of the issues. As indicated above, a century ago it was natural for scientists to focus on the great diversity among individual patients and the complex interactive character of the available therapies. The advent of powerful antibiotics changed all that. Therapies became highly standardized, primarily in the form of drugs, and the doctor’s role was limited to diagnosing the disease and prescribing the correct medication and dose. Patient outcomes became standardized as well; either the patient recovered or failed to recover.

Specific underlying factors that affected the outcome, however complex they might be, were completely invisible. For all practical purposes, the outcomes were like rolls of the dice or spins of the roulette wheel. Thus, the conditions were ideal for applying the same experimental methods that had proved so useful in agriculture and industry. Individuals might vary, but the underlying causes of variation were inaccessible and essentially irrelevant. The increase in the rate of cure in the population could be assumed to provide important information about any individual’s increased “chances” of being cured. So, what is wrong with this logic?

The problem is not with the methodology per se, but with its fit to current conditions. Increasingly, the chronic diseases of aging that are being tackled by pharmaceutical R&D today do not conform to the classic statistical assumptions. In fact, they raise more sophisticated versions of the same issues that vexed researchers long ago. These chronic conditions have come into focus precisely because of improvements in health and living standards made possible by science, especially medical science. But afflictions like heart disease, cancer, and Alzheimer’s have much more complex and multi-factorial etiologies than bacterial infections. Effective weapons against these chronic conditions take the

form of long-term lifestyle modification along with medications that target various biological processes. These new drugs essentially interact in complex ways with individual biology and often with other drugs as well. Treatment becomes a complex balancing act, as the physician tinkers with delicate biophysical and biochemical mechanisms that are only partially understood.

These changes in the nature of primary target diseases and their treatment have major implications for the design of research. Once again, the physician is faced with the salience of individual variability. Dispensing medications in a one-size-fits-all manner does not make sense. Helping each individual patient navigate the complex range of options for optimizing her health is a fundamentally different problem than curing a case of syphilis or pneumonia. The kind of information about “average” effects typically obtained from RCTs must be filtered through the doctor’s experience to determine its possible relevance to the individual patient. But most reports of RCTs offer little or no meaningful guidance for a “personalized” interpretation of efficacy and safety. Moreover, the rapidly expanding knowledge being generated about genomics and biomarkers will only exacerbate the disconnect between RCT results and clinical practice. How should the results of sophisticated genetic and other testing be entered into the equation when treating each individual?

## 6. The Great Misconception

The randomized clinical trial is regarded by most statisticians as the only way to estimate the “true” effect of a medical intervention. Despite the growing importance of personalized medicine, overall (population-level) effects are considered to be the first-order lessons to be learned from any RCT. Only after such a general effect is established, can the “interactions” between the treatment and other variables be considered. However, these “second-order” interactive effects, if they exist, are much harder to prove (lower statistical power, problems of multiplicity) and are accordingly treated as tentative, requiring further investigation (which rarely occurs). Thus, the cards are stacked against detecting any characteristics that can either enhance or dilute the drug’s effect.

In the 1950’s, it was perfectly reasonable to view the “main effect” estimated from a properly designed RCT to be the primary issue of clinical importance. This population-level effect could be interpreted as a measure of how much *expected* benefit in terms of decreased “risk” a patient might obtain. But increasingly the belief that an aggregate effect for a population is relevant to each individual may be a misconception. Perhaps we need to recover some of the skepticism of our nineteenth-century predecessors regarding such statistical summarizations of disparate “facts,” albeit from a more sophisticated modern vantage point.

To clarify the issues, let us consider the simplest situation, when the outcome of interest is a particular event (say an ischemic stroke). An RCT comparing a proposed new preventive treatment typically will be summarized by a relative risk, perhaps expressed as a risk ratio or hazard ratio. This global parameter, if found to be statistically significant in a properly conducted RCT is interpreted as the estimated causal effect on the outcome. Such a finding is the primary green light that is necessary for approval by regulatory agencies. What is rarely understood clearly, however, is that this effect is a statistical summary of the *individual effects* for the particular *population* under study.

What do we mean by an individual causal effect? The classical probabilistic model underlying modern statistics has no way to represent the idea of an individual effect. However, in recent decades methodologists have recognized that we can define, in theory, the *potential outcomes* that would occur under two different treatment modalities. For example, the “outcome” for each subject, such as whether or not a stroke occurs, is conceived as a pair of possibilities, a *response pattern*. The problem is that only one of the two possibilities is actually observed. The other is “missing.” By calculating a global effect parameter, we effectively “average” the unobservable individual effects.

In an early paper on agricultural experimentation published in Poland in 1923, Neyman explicitly defined this counterfactual causal effect for each experimental unit. However, neither he nor anyone else pursued the implications of this idea. About 50 years later, Donald Rubin introduced the same concept, initially as a way to express and tackle problems of bias in observational studies. Building on this idea, Rubin and his colleagues have derived such important innovations as propensity scores and principal components. However, while recognizing the existence of individual effects, these methods have so far remained primarily focused on aggregate summary measures, such as the average effect or relative risk. But suppose that these effects are highly variable, as may well often be the case in the kinds of research of most interest today. Then what does this aggregate effect really mean?

## 7. Ambiguity

Modern statistical methods have achieved great success in many areas of application. However, these technical triumphs depend on the suppression of effect variability. When causal effects can vary substantially across individuals, there is an unavoidable *ambiguity* in the interpretation of aggregate effects. This ambiguity arises because there are many possible distributions of individual effects that are consistent with any particular aggregate effect. If the individual effects are approximately uniform or vary in an essentially random manner (i.e. unrelated to any potentially observable factors), then the population parameter remains relevant to any member of the population. However, suppose there exist potentially identifiable individual characteristics that are related to the causal effect. Then it may become feasible to specify subgroups based on these characteristics for which the overall parameter value is misleading.

Taking seriously this potential leads to complexities that are beyond the power of “pure mathematics” to resolve. Neyman, at the dawn of modern statistics, was perhaps uncomfortable with the resulting ambiguities, but may have reasoned that in the context of agricultural experiments, the variability across plots of land was immaterial to the paramount issue of the overall yield. In modern clinical research, this approach is becoming increasingly untenable. Consider the example of a hypothetical drug to prevent stroke mentioned above. There are four possible response patterns for any subject, as shown in Table 1.

Table 1			
Four Possible Response Patterns			
Response Pattern	Treatment	Placebo	Proportion
1: Doomed	stroke	stroke	$P_1$
2: Causal	stroke	no stroke	$P_2$
3: Preventive	no stroke	stroke	$P_3$
4: Immune	no stroke	no stroke	$P_4$

The expected causal effect calculated from such a study is a function of the distribution of response patterns. For example, the relative risk is given by:

$$RR = (P_1 + P_2) / (P_1 + P_3)$$

To understand why, observe that an observed stroke in the study group that is given active treatment can represent either a *doomed* or *causal* individual. Likewise, a stroke in the comparison group can be either a *doomed* or *preventive*.

The same value of  $RR$  can correspond to many possible underlying distributions of the response patterns. Suppose, for instance, that  $RR = 0$ . One possible underlying distribution of response patterns is shown in Table 2.

Table 2			
Example of Sharp-Null Hypothesis			
Response Pattern	Treatment	Placebo	Proportion
1: Doomed	stroke	stroke	.20
2: Causal	stroke	no stroke	0
3: Preventive	no stroke	stroke	0
4: Immune	no stroke	no stroke	.80

This underlying causal structure would generate the following observed 2 x 2 table:

Table 3			
Hypothetical Stroke Study Results			
	Stroke	No Stroke	
Treatment	200	800	20%
Placebo	200	800	20%

These results would ordinarily be interpreted to mean that the intervention had absolutely no impact. This notion that no effect for any individual is sometimes called the “sharp-null hypothesis.” However, Table 3 would also be consistent with the underlying structure displayed in Table 4:

Table 4			
Example of Dull-Null Hypothesis			
Response Pattern	Treatment	Placebo	Proportion
1: Doomed	stroke	stroke	.10
2: Causal	stroke	no stroke	.10
3: Preventive	no stroke	stroke	.10
4: Immune	no stroke	no stroke	.70

This “dull-null” hypothesis would tell quite a different story, *if the underlying causal structure could be revealed*. Rather than an overall risk ratio, researchers would focus on understanding exactly who could be helped and who harmed by the drug. Can we specify variables that predict treatment success? Can we find subgroups of patients with distinct distributions of response patterns that differ from the overall population distribution? These questions are rarely asked, especially if the overall effect parameter is not statistically significant.

This causal modeling perspective also has important implications for the design of RCTs. Sample sizes are chosen to achieve a desired level of statistical power. For example, a study might be powered to have a .80 probability of declaring a significant effect if the true  $RR=1.50$ . However, just as for the null hypothesis, this alternative hypothesis is also ambiguous. We would interpret this  $RR$  to mean that a stroke would be avoided for half of those subjects destined to experience one without treatment. We would thus implicitly assume the situation portrayed by Table 5.

Table 5			
Usual Alternative Hypothesis for $RR = 2$			
Response Pattern	Treatment	Placebo	Proportion
1: Doomed	stroke	stroke	.10
2: Causal	stroke	no stroke	0
3: Preventive	no stroke	stroke	.10
4: Immune	no stroke	no stroke	.80

Power calculations are based on the probability distribution for two independent binomial samples, regardless of the underlying causal structure. However, it is possible that a subgroup of the subjects would actually be harmed by the treatment. Thus, the true state of affairs might be as shown in Table 6.

Table 6			
Another Alternative Hypothesis for $RR = 2$			
Response Pattern	Treatment	Placebo	Proportion
1: Doomed	stroke	stroke	.05
2: Causal	stroke	no stroke	.05
3: Preventive	no stroke	stroke	.15
4: Immune	no stroke	no stroke	.75

There are an infinite number of possible alternative hypotheses that all correspond to  $RR = .50$ . So, calculating power against a global alternative like  $RR = .50$  could be irrelevant if our real interest is in stratifying the population for potential personalized treatment. The many possibilities for post hoc analyses make a shambles of any traditional measurement of statistical significance or power.

The existence of individual effect variability is an inconvenient truth for standard RCT methodology. So, it is convenient for the *statistician* to ignore it. However, the *clinician* may have reasons to believe that her particular patient has special characteristics that are relevant to his chances of responding favorably to the treatment. Such qualitative insight effectively places the patient in a different “reference population” from that on which the RCT was based. In the nineteenth century, the physician’s judgment would have been trusted much more than the “numerical” rates derived from a large anonymous group. Today, we have come full circle, so that this hard-to-quantify clinical expertise can be given virtually no weight in our analyses. Has this pendulum swung too far?

## 8. The Future of Clinical Trials

Clinical trials are typically designed with exquisite care around a single objective: to guarantee that the primary null hypothesis can be proved or disproved. The protocol for an RCT specifies the number of subjects necessary to assure adequate statistical power and the exact statistical techniques to be employed. These *procedural* steps provide the sole rationale for believing the results of the study. However, the iron discipline of the modern RCT also exerts a highly conservative influence. Investigators must effectively promise to learn absolutely nothing throughout the long years of research. They are not only “blinded” but effectively lobotomized. As a result, the researcher becomes a passive participant who is unable to think creatively about what is happening and to refine his understanding as new data and external information emerge over time.

The sacrifices made by investigators and patients to participate in RCTs are usually justified on the basis of necessity. In order to be *certain* about the one pre-specified piece of information, any deviations from the study plan must be avoided. But in light of individual effect variability, such certainty is illusory. As a result, the payoff for the enormous investment of time and effort is often much too paltry. Consider that the majority of new medical products fail to gain approval after years of study. Most are shelved, never to make a contribution to human health. Those that are ultimately approved may later be found to have unforeseen adverse effects that limit their use. The

true calculus of risk versus benefit for the drug or device may turn out to be variable across different populations. This problem of external validity has been largely ignored because the potential importance of effect variability has not been appreciated.

Rather than narrowly focusing on whether or not the treatment “works” in general, we should ask a better question. *For whom* (if anyone) is the treatment beneficial and *for whom* is it harmful? What individual and circumstantial characteristics are conducive to a positive (or negative) response? To answer such questions will require a more flexible approach to design and analysis of RCTs. There needs to be some mechanism for extracting information about effect variability and its possible implications during the course of the trial. For example, some researchers could be encouraged to commit the heresy of formulating new or modified hypotheses and to perform post hoc interim exploratory analyses. This “discovery team” would be completely separate from the more traditional research infrastructure. In effect, its mission would be to “mine” the data, both during and after the trial.

Incorporating a discovery component would greatly increase the potential value of the RCT. The additional knowledge gained could set the stage for a more nuanced presentation of study findings. It might suggest, and to some extent test, the importance of various causal factors related to genetics, concomitant medication, medical history, etc. This information could aid clinicians in personalizing their treatment decisions. Even if the results of the data mining efforts are negative, they could prove useful by disabusing practitioners of intuitions that may be incorrect. Of course, the exploratory research would be less “rigorous” than the traditional form of evaluation. As an exercise in data mining, the strength of evidence produced would be assessed not by statistical significance, which would be virtually impossible to assess, but would depend mainly on *validation*, either within the RCT itself or in subsequent research.

In the burgeoning field of “predictive analytics,” data mining techniques, including independent validation on hold-out samples, has become standard practice. The philosophy is to “trust but verify” under the optimistic assumption that meaningful “segmentation” of a population is likely to exist. Techniques that explicitly aim to detect characteristics associated with different response patterns for treatments are starting to be developed, primarily in the context of marketing research. Finding productive ways to bring this sort of creative energy into the sphere of clinical trials without sacrificing scientific rigor is a critical challenge for today’s statisticians.

## References

Bernard, C. (1865). *An Introduction to the Study of Experimental Medicine* (Translation published 1957 ed.). Mineola, NY: Dover Publications.

Chen, T. T. (2003). *History of Statistical Thinking in Medicine*. In *Advanced Medical Statistics* (Chapter 1, pp. 3-19). Singapore: World Scientific Publishing.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Greenland, S., & Robins, J. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology* , 15, 413-19.

Guy, W. A. (1860). Croonian Lectures on the numerical method, and its application to the science and art of medicine. *British Medical Journal* .

Huth, E. J. (2006). Elisha Bartlett (1804-1855), an American disciple of Jules Gavarret. *JLL Bulletin: Commentaries on the History of Treatment Evaluation* , (www.jameslindlibrary.org).

Huth, E. J. (2006). Jules Gavarret's *Principes Generaux de Statistique Medicale*: a pioneering text on the statistical analysis of the results of treatments. *JLL Bulletin: Commentaries on the History of Treatment Evaluation* , (www.jameslindlibrary.org).

Laplace, P.-S. (1825). *A Philosophical Essay on Probabilities* (Fifth ed.). Paris: Bachelier.

Lewis, S. (1925). *Arrowsmith*. New York: Collier.

Lo, V. S. Y. (2002). The true lift model: a novel data mining approach to response modeling in database marketing. *SIGKDD Explorations*, 4, 78-86.

Lowy, I. (2010). Martin Arrowsmith's clinical trial: scientific precision and heroic medicine. *JLL Bulletin: Commentaries on the History of Treatment Evaluation* , (www.jameslindlibrary.org).

Marks, H. M. (1997). *The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900-1990*. New York: Cambridge University Press.

Mill, J. S. (1874). *A System of Logic* (Eighth ed.). New York: Harper and Brothers.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: essay on principles. *Annals of Agricultural Science* .

Penston, J. (2005). Large-scale randomised trials: a misguided approach to clinical research. *Medical Hypotheses* , 64, 651-57.

Poisson, S. D., Dulong, P. L., Larrey, D. J., & Double, F. J. (2001). Statistical research on conditions caused by calculi by Doctor Civiale (with commentaries). *International Journal of Epidemiology* , 1246-1258.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* , 66, 688-701.

Susser, M. (1977). Judgment and causal inference: criteria in epidemiologic studies. *American Journal of Epidemiology* , 105, 1-15.

Weisberg, H. I. (2010). *Bias and Causation: Models and Judgment for Valid Comparisons*. Hoboken, NJ: John Wiley & Sons.