The Evolution of Statistics in Medicine

Past, Present, and Future

Independent Study

Statistics in Medicine

Kate Brieger

December 2010

Table of Contents

"It is not the business of the Mathematician to dispute whether quantities do in fact ever vary in the manner that is supposed, but only whether the notion of their doing so be intelligible; which being allowed, he has a right to take it for granted, and then see what deductions he can make from that supposition" (Barnard & Bayes 1958).

Introduction

In the United States, more than twice as much is spent per person on healthcare as in most other industrialized nations. Despite soaring healthcare coast, the country is failing, comparatively, at preventing deaths through the use of effective and timely medicine. There is a high dependence on individual physician judgment which often leads to heroic measures, including expensive and futile treatment. Statistically sound studies are often ignored and there is a lack of application of recommended public health interventions. Evidence-based medicine is, in fact, quite a new paradigm in the world of healthcare. The medical community supports the use of current medical literature and the results of clinical trials to determine the best course of treatment for a given patient. Although physicians think of themselves as far-removed from the pre-historic men who believed that illness was an entirely spiritual event, throughout most of the 20th century there was no use of statistics to evaluate the effectiveness of new medical technologies. Advances were made mostly through study of physiology and physicians frequently used individual case studies to "prove" their theories.

Finally, in the 1950's, the randomized clinical trial became the new standard for research. The history of modern statistics in medicine is surprisingly short. Just decades ago, medical studies did not use control groups, placebos or large sample sizes. In recent years, statistical methods have rapidly been adapted for the description and analysis of medical issues. Still, though, many statistical tests and summaries remain misunderstood and inadequately presented in the current medical literature.

Public Perception

Opposition to the use of statistical studies by physicians may be traced to competing views of the physician as an artist, determinist, or statistician (Senn 2003). In 19th century Paris, the Royal Academy of Medicine adopted the beliefs of Risueno d'Amador that the use of statistics was inadvisable as relying on statistics would not cure "this or that disease, but . . . the most possible out of a certain number," in effect "condemning" certain individual patients to death. D'Amador concluded that physicians, like artists, should rely on intuition to sculpt each unique patient. A determinist physician, on the other hand, is one who relies on experimentation and believes that they can treat with certainty; determinists emphasize that science looks to find causes and not chances, playing to the patients' desire for predictable outcomes. On a fundamental level, the advance of statistics may be held back by the human desire for certainty, even where none exists.

The general population holds a common mistrust, and even contempt, of statistics. People recognize that numbers are often misleading, perhaps citing the example of one billionaire dramatically altering the mean income of a city's residents. Health statistics are no less convoluted, and arguably much more so. For instance, the literature says that mesothelioma is incurable and has a median mortality of eight months after diagnosis (Gould 1985). The reality, of course, is that there is variation and that means and medians are abstractions. Gould, an optimistic patient, noticed that the variation about the eight-month median was right skewed, with a very long tail; he went to live 20 years after his diagnosis. "So far as Mathematics do not tend to make men more sober and

rational thinkers, wiser and better men, they are only to be considered as an amusement, which ought not to take us off from serious business" (Barnard & Bayes 1958). Many people continue to regard statistics as superfluous information that cannot completely describe any clinically relevant results or "serious business."

Further, even leading medical journals present nontransparent statistics. People, including physicians, are then less likely to make the effort to understand the results beyond what the author summarizes. For example, relative risks shown without their corresponding base rates are misleadingly large. Another source of confusion is when benefits and harms are reported using different measures, as when relative risk reduction suggests a large benefit and absolute risk increases suggest a small harm. Because the media dramatizes these numbers, the public can become unnecessarily alarmed or in appropriately comforted.

History of Clinical Trials

Before 1750, there were no clinical trials. Following the ancient Greek model of the cause of disease, physicians treated patients with the goal of restoring balances in blood, phlegm, and bile (Green, Benedetti & Crowley 2002). Cancer, for instance, was treated with rigorous purging and a bland diet to avoid the congestion of black bile. Finally, in the 1800s, numerical methods came into favor. Duvillard (1806) showed with a primitive analysis that smallpox vaccination decreased the general mortality rate. By the end of the 19th century, the principles of comparative trials had been described (Bernard, 1866) and even suggested as a remedy for the "doctor [that] walks at random and becomes sport of illusion." Sir Arthur Bradford Hill wrote a series of statistics papers for *The Lancet* that was published as a book in 1937, arguing for randomized clinical trials. Hill was familiar with the idea of randomization from the work of R. A. Fisher, a noted scientist in the design of agricultural experiments (Hill 1937). Fisher supported the practice of running experiments with concurrent control groups, as opposed to making historical comparisons (Senn 2003).

Unfortunately, it was not until 1946 that the first randomized therapeutic clinical trial was conducted. Given that there was a limited supply of streptomycin, the proposed treatment for tuberculosis, Hill argued that a strictly controlled trial was necessary. In a sense, this trial began a new age of evidence-based medicine. In 1954, the largest medical experiment in history was carried out with over a million children to test the effectiveness of the Salk vaccine in protecting against poliomyelitis (Meier 1977). The study used a placebo control, assigned treatment groups randomly, and evaluated

outcomes using a double-blind model. Polio was a serious disease that came in epidemic waves and left many cripples; in addition, President Franklin D. Roosevelt supported the search for a vaccination after contracting the disease himself. It would have been simple to distribute the Salk vaccine as widely as possible, but this would have failed to produce clear evidence because polio varies annually and geographically. Also, since the diagnostic process is influenced by the physician's expectations, leading to the necessity of the double-blind design. Finally, the control group was necessary because the families that would volunteer children to receive vaccination are inherently different than those that would not.

Today, it is clear that randomized trials are necessary to test new drugs, but the trials must be developed with more clearly defined and statistically relevant stopping criteria. Ethically, the new drugs must be compared to existing treatments instead of placebo because sick patients cannot be denied treatment. The control group is still useful, though, since what the researcher truly wants to discover is whether the new treatment is better than the current protocol. The situation is more complicated if there is no existing drug and the new drug is very promising. Without randomized trials, efficacy is impossible to prove; further, there have been many disasters when drugs were brought prematurely out of the trial phase because of pressure and off label use. On the other hand, when the choice is between receiving no treatment and receiving an experimental one and the outcome is fatal without intervention, it is difficult to argue to continue randomized trials and deny patients a potentially lifesaving intervention. Perhaps the future holds better testing in animals or other simulations to enable quicker movement of

drugs through the process. However, there will remain inherent differences between any model system and real humans.

Ethics of Clinical Trials

Some ethicists find randomization to be a repugnant practice because patients in a clinical trial are knowingly subjected to treatments with incompletely understood effects. To minimize the unethical aspects of trials, researchers must be perfectly indifferent as to which treatment is better. Once the trial reaches a point when the investigators believe that one treatment is better, the investigators cannot continue to randomize or else they are "sacrificing the interests of current patients to those of future patients [and] to treat patients as means and not ends and is unethical" (Senn 2003). A trial can legitimately continue until either the investigator is convinced that one treatment is more efficacious or convinced that there is not a difference.

There are additional complicating factors involved in randomization. If a patient does not have insurance, they want to participate in the study even if assigned to the control group. When a new drug is promising in early tests, some argue that the trial delays access to the drug and causes needless suffering. Others, though, point out the problems with not using a control group because researchers should be required to prove that they are helping people in the long run and that the new drug prolongs life expectancy. When sick people assigned to the control arm ask to switch to the "treatment" arm, they are denied; some argue that scientists already know what the outcome of the trial will be, but are leaving people on the control arm because they need them to die earlier to prove a point.

Even before data is collected, controlled clinical trials have important statistical components. Statistics is used to determine the randomization, blocking, sample size,

and power. Perhaps one of the most difficult criteria to set is the stopping rule because there is an ethical conflict between trying to ensure that the study participants receive beneficial treatment and that the competing treatments are effectively evaluated for future patients. Phase II trials aim to investigate drug efficacy while still monitoring toxicity (Nguyen 2009). A phase II trial should be halted in there is sufficient information to make a conclusion or if a large proportion of patients experience toxicity effects. There is ongoing research and discussion about the best way to set the stopping criteria. One issue is that the continuous monitoring, after each patient enrollment, inflates the type I error because the investigators running the trial are doing a sequential test where there is a chance of making a type I error after each patient. Any method used for determining a stopping boundary assumes that investigators continue testing as long as there is insufficient evidence to stop the test.

Improperly designed experiments are unethical to carry out because they will not provide useful information and are therefore a waste of time, energy, and human subjects. Since statistical methods are one aspect of experimental design, they deserve emphasis at every stage of the trial. Experiments with too few subjects for valid results or an improperly designed random or double-blind procedure are a serious breach of ethics (Altman 1980). Errors in analysis and interpretation of results can be rectified before publication, but deficiencies in design are irremediable. Many possible biases in analytic results can occur in planning, design, data collection, data processing, data analysis, presentation, interpretation, and publication (Sackett 1979). No matter in which stage of the process the error occurred, it is unethical to knowingly publish results lacking in statistical integrity.

Errors in Clinical Trials

There are countless examples of studies being misreported. The general public, without adequate statistics background, erroneously applies the information that they learn through the media to their lives; people become unnecessarily wary of existing treatments or unfairly excited about new treatments. For instance, a Viagra special report linked the sex drug to 31 deaths in one year and caused great concern (Viagra Special Report 2000). What was missing, however, was an estimate of the total exposure to Viagra and if the number of deaths in the population of Viagra users was greater than expected, given their numbers, their age, and the time they were taking the drug.

The effects of bad reporting can have serious public health implications. The spring of 2002 saw a panic about the vaccine for measles, mumps, and rubella (MMR). The dangers of the vaccine made headlines in the United Kingdom, warning parents about the vaccination's associated risks for autism and inflammatory bowel disease. The alarm was based on a 1998 paper from *The Lancet* that reported a study on 12 children who had gastrointestinal disease and developmental regression (Wakefield et al. 1998). The parents of 8 of the 12 children associated the onset of the health problems with their children having been given MMR. The statistics for the general population, though, seem to suggest that the alarm was unfounded. Based on the World Health Organization's figures about immunization and autism rates, finding 12 children who had received MMR and also had autism is not remarkable. In fact, if none of the children had received MMR, it would actually have indicated that MMR protected against autism. Additionally, the symptoms of autism are often first noticed at the same age as when

children receive vaccination, so the association in the parents' minds may have been coincidental. However, researchers still cannot be sure that MMR does not cause autism; it is nearly impossible to prove that something is safe. In general, vaccination can be seen as a public health issue with externalities, implying that political intervention may be necessary to provide the greatest good for the greatest numbers. The implementation of policy, though, is far from simple. One solution could be to offer health insurance cost reduction to those who vaccinate their children.

In a similar vein, the saga of hormone replacement therapy had widespread effects. By the early 1990s, numerous observational studies had found lower rates of coronary heart disease (CHD) in postmenopausal women who took estrogen than in women who did not. However, the potential benefit of hormone therapy had not been confirmed in clinical trials. The objective of the HERS trial was to determine if estrogen plus progestin therapy altered the risk for cardiac events in postmenopausal women with coronary disease (Hulley et al. 1998). The randomized, blinded, placebo-controlled study was conducted at 20 U.S. clinical centers with a total of 2783 women. The results indicated that there were no significant differences between groups in the primary outcome or in any of the secondary cardiovascular outcomes. There was, however, a statistically significant time trend where more CHD events occurred in the hormone group than in the placebo group in year 1 and fewer in years 4 and 5. Further, more women in the hormone group than in the placebo group experienced venous thromboembolic events and gallbladder disease. The study concluded that there was no overall cardiovascular benefit and that there was a pattern of early increase in risk of

CHD events; therefore, the researchers did not recommend starting this treatment for the purpose of secondary prevention of CHD.

When the HERS findings were published in *JAMA* in 1998, the prevailing reaction was disbelief and the results were largely ignored. At the time, Premarin was the most widely prescribed drug in the United States. The drug's popularity was partly based on its historic role in the treatment of menopause symptoms, as it had been approved in 1942 by the FDA for the treatment of hot flashes. The well-read book *Feminine Forever* popularized the philosophy that menopause is completely preventable because the condition was a simple hormone deficiency (Wilson 1966). The book was written by a physician, but was misleading and immodest. Additionally, animal studies suggested that estrogen could slow the rate of atherogenesis and small-scale trials found that hormone treatment increased high-density lipoprotein ("good") cholesterol and improved endothelial function.

With the conclusion of the Women's Health Initiative study, the findings of the HERS trial were supported. The trial had a similar design as HERS but used a much larger sample size (16,608) and used women free of coronary heart disease. Hormone therapy significantly increased rates of CHD, stroke, pulmonary embolism, and breast cancer. Despite the low absolute magnitude of the increased risks, the harms are substantial given that the treatment is design for healthy women. Practice guidelines now recommend that hormone therapy be used at the lowest possible dose and for the shortest possible time. Finally, in 2002, the number of hormone prescriptions decreased (Hersh et al. 2004). The decrease in hormone therapy has been associated with a decreased incidence of estrogen receptor-positive breast cancer (Jemal et al. 2007). Evidence-based

medicine is the new paradigm that practice guidelines must be based on rigorous research, keeping in mind that animal studies and epidemiologic studies are often misleading. Accurately analyzing the benefits and harms is particularly crucial in the consideration of preventive interventions for healthy individuals.

Publishing Results

One of the most serious problems in medical research is that biological understanding and previous research play little formal role in the interpretation of quantitative results. There is the illusion that conclusions can be produced with certain error rates and that their importance can be inferred from their significance level. Hypothesis tests are pure statistical approaches, but people are mistaken if they believe that a single number, the p-value, can capture both the long run outcomes of an experiment and the evidential meaning of a single result. Because it is rare that studies examine issues about which nothing is already known, they must always be interpreted in the context of the field. Compared to hypothesis testing, the Bayes factor is more comprehensive because it properly separate issues of long run behavior from evidential strength and allows integration of background knowledge with statistical findings (Goodman 1999). "It is commonly believed that anyone who tabulates numbers is a statistician. This is like believing that anyone who owns a scalpel is a surgeon" (Hooke 1983). Not everyone understands the nuances of different tests and so not everyone is qualified to interpret the published results.

On a more basic level, medical literature shows a strong tendency to accentuate the positive. Because positive results are more likely to be reported (Berlin, Begg & Louis 1989), some purely chance findings will be published and mistakenly be considered important. Traditionally, journals set the standard of a p-value smaller than 0.05 to provide strong evidence against the null hypothesis, but this creates an arbitrary division of results into "significant" or "non-significant." This division was not among

the intentions of the founders of statistical influence. Precise p-values should be presented without reference to arbitrary thresholds and results of medical research should be interpreted in the context of the type of study and other available evidence. Fisher himself argued that interpretation of the p-value was for the researcher, helping him determine whether to perform another experiment. Along those lines, the Neyman-Pearson approach requires that the scientist specify a precise null hypothesis and makes no attempt to interpret the p-value to assess the strength of evidence against the null hypothesis in an individual study (Neyman & Pearson 1933).

Further, researchers have cognitive biases, doctors and patients have emotional relationships, and the healthcare system has conflicts of interest. Information pamphlets and websites produced by pharmaceutical companies tend to suggest that the newly featured intervention offers great benefit and little harm. Because the numbers are reported in a confusing way, the ideal of shared decision-making and informed consent is all but abandoned.

Clinical Applications

Physicians are not expert in the interpretation of new studies and so patients are not given cutting-edge, evidence-based treatment. In 2000, a grand round survey found that physicians were better in basic numeracy that the general public, but still not fully competent. Hoffrage and Gigerenzer (1998) determined that physicians are often confused by sensitivities and specificities, making it difficult for them to give patients helpful advice about screenings. Cancer screening is a prime example of an innovation that has caused an increase in survival rates without prolonging life. Suspiciously high survival rates may be due to the overdiagnosis bias, where screening detects abnormalities that will never progress to cause cancer symptoms. For instance, with computed tomography scanning, cancer was detected in nearly as many nonsmokers as in smokers (Sone et al. 2001). Knowing that 15 times as many smokers die of lung cancer, it is clear that the scans were picking up abnormalities that would not behave like a life threatening metastasizing cancer. Higher survival rates after the increased prevalence of cancer screening does not mean that patients are living longer. Mortality, not survival rates, is what patients should be interested in.

Mammography also has a high rate of false positives; in fact, 9% of women without breast cancer will test positive and become unnecessarily concerned. Reporting the conditional probability ("if a woman has breast cancer, the probability that she tests positive is 90%") is more confusing than reporting the natural probability ("of the 10 women in 1000 who have breast cancer, 9 will test positive"). Natural frequencies are easier for most people to understand than conditional probability because humans

automatically encode numerical information this way, without having to be taught probability concepts.

Besides the complicated results of screening tests, regional customs outweigh evidence for many types of treatment because physicians simply cannot interpret new findings. In Maine, for instance, the proportion of women who have undergone a hysterectomy ranges between regions from less than 20% to more than 70%. Similarly, in 1995 in the United Kingdom, it was announced that a new oral contraceptive pill increased the risk of life-threatening blood clots by 100% compared to the previous version of the pill. The warning issued by the Committee on Safety of Medicines strongly influenced women's decisions; many stopped taking the pill and the scare led to an estimated 13,000 additional abortion in the following year. The warning, it turns out, was based on the fact that the new pill caused thrombosis in 2 out of 7,000 women taking the pill as opposed to 1 out of every 7,000. That is, the absolute risk increase was only 1 in 7,000. In general, absolute risks are small while the relative changes look large, particularly when the base rate is low. Most pathological conditions are rare and so the relative changes when comparing treatments are often misleadingly impressive (Furedi 1999).

Finally, the process of dying has become implicated in statistics because of how doctors try to predict and prevent death. Before the age of modern medicine, dying was a brief process that we did not pour substantial resources into prolonging. When cancer patients go to see social workers, they do not want to focus on survival statistics; they want to focus on aggressive, pull-out-all-the-stops treatment. In the United States, the astronomical cost of health care is in large part due to the terminally ill; 25% of Medicare

spending is on the 5% of patients in their last year of life. For cancer survivors, the majority of the cost is for initial diagnostic testing, surgery, radiation, and chemotherapy. For patients with fatal breast cancer, the average cost of the last six months of life is $63,000 (Gawande 2010).

Interestingly enough, it appears that a patient's life expectancy does not change when they stop trying to fight death. Medicare patients in hospice and those not in hospice showed no difference in survival time, despite the fact that hospice does not use invasive and expensive procedures simply to extend life. Even doctors have unrealistic views of how much longer their patients can live. Christakis (2003) found that when doctors were asked to predict how long their terminally ill patients would live, the average estimate was 530% too high. Clearly, emotional involvement can play a greater role than statistical evidence in patient-doctor interactions.

Statistical Education

Health statistics not only confuse patients, but also many doctors, journalists, and politicians. The rampant "statistical illiteracy" that persists in our society is in large part due to the nontransparent framing of information, leading to serious consequences that prevent optimal treatment (Gigerenzer et al. 2008). The problems with statistical education are by no means being realized for the first time today. In 1937, an article in *The Lancet* criticized physicians' "blind spot" in laboratory and clinical medicine. The editorial claimed that physicians ended up woefully unprepared because "simple statistical methods concern us far more closely than many of the things we are forced to learn in the six long years of the medical curriculum." Further, people did not realize that much of the data being generated was statistically inadequate and lead to avoidable errors and "a sad waste of materials." A decade late, in 1948, the British Medical Association recommended that statistics be included in medical education, but it was not until 1975 that statistics became mandatory at the University of London (Altman & Bland 1991).

A survey completed by nearly 300 residents in 11 different programs confirmed that there is only minimal understanding of statistical concepts among residents (Windish, Huot & Green 2007). The mean percentage correct on the test of statistical knowledge and interpretation of results was 41.4% for residents; among fellows and general medicine faculty, however, the mean score was 71.5%. Of those surveyed, 75% admitted that they did not understand all of the statistical concepts they came across in journals, yet 95% reported that they believed it was important to understand these concepts to be "an intelligent reader of the literature." On the 20-question test, residents

scored highest in recognition of double-blind studies and in interpretation of relative risk; in contrast, very few were able to interpret odds ratios. The number of years since completing medical school was correlated with a decrease in score, indicating that what little statistical education is provided during medical school is not retained.

To provide patients with the best available care, it is imperative that physicians correctly interpret results published in the most recent journal articles. Clearly the statistical instruction provided to residents must increase in quantity and quality, to ensure that they are being trained to become responsible, up-to-date physicians. The current education system emphasizes "mathematics of certainty," such as algebra and geometry. Statistical thinking is usually introduced late in school, with confusing representation. Statistical thinking should be taught in medical school, but also early on as part of primary and secondary education. Incorrect interpretation of results can lead to inappropriate and even dangerous applications of clinical research. In the 1990s, a comprehensive survey of medical school biostatistics teaching found that an overwhelming majority (more than 90%) focused their biostatistics teaching in the preclinical years without any subsequent instruction (Looney et al., 1998).

Into the Future

Statistics will maintain its place as a critical analysis tool, particularly as the theory and methods are refined and computer programs are developed. For centuries, certain scholars have recognized the power and beauty of statistics. "Some people hate the very name of statistics but I find them full of beauty and interest. Whenever they are not brutalized, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary. They are the only told by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of man." (Galton 1889). Currently, the public, and particularly those involved in healthcare, should aim to better appreciate the crucial nature of statistical analyses.

Technological advances are rapidly changing the role of statistics. Researchers are now capable of building enormous data bases through methodical data collection. What is more, they have access to computer power and international cooperation to carry out the statistical investigations (Schoonjans et al. 1995). Some computer scientists developing artificial intelligence and machine learning are further developing Bayesian methods, while others are working to refine techniques of data-mining and bioinformatics. Statistical theory and algorithms are rapidly improving with advances in technology (Collen 1992).

However, much remains to be explored. For instance, the standard two-arm design in clinical trials has a clear role, but the advantages and disadvantages of multi-arm designs remain poorly understood. As methods become more advanced, it is

increasingly important to consider type II error rate. In medical journals, there has been progress in the transparency of reporting. Now, confidence intervals are presented in addition to p-values (Gardner & Altman 1986), helping to mitigate the problem of the common misinterpretation that the p-value is the probability that the null hypothesis is true. The most important improvements to the quality of clinical trials can come from increasing sample size and precision of measurement. Journal editors must take responsibility to assure that results are reported in an intelligible manner. Further, results should always be reported with confidence intervals and should not be stated as simply "significant" or "not significant" (Sterne & Smith 2001). For researchers to best take advantage of advances in statistical methods, the readers of journals must also become statistically literate. Over the coming years, health statistics can lead to improvements in medical practice that will save countless lives.

References

Altman DG & Bland JM. 1991. Improving doctors' understanding of statistics. *Journal of the Royal Statistical Society*. Series A. 154, 223-267.

Altman DG. 1980. Statistics and ethics in medical research. Misuse of statistics is unethical. *British Medical Journal.* 281:1182-1184.

Barnard GA & Bayes T. 1958. Studies in the History of Probability and Statistics: IX. Thomas Bayes's Essay Towards Solving a Problem in the Doctrine of Chances. *Biometrika*.

Berlin JA, Begg CB, Louis TA. An assessment of reporting bias using a sample of published clinical trials. *Journal of the American Statistical Association*. 1989. 84:381-92.

Bernard CL. (1866). Introduction à l'Etude de la Médecine Expérimentale. Granier-Flammarion: London.

Christakis N, Simes J, Glare P, Virik K, Jones M, Hudson M, Eychmuller S. 2003. A systematic review of physicians' survival predictions in terminally ill cancer patients. *British Medical Journal.* 327;195.

Collen M, Bland FLM & Altman DG. 1992. Comparing Two Methods of Clinical Measurement: A Personal History. *A history of medical informatics in the United States, 1950 to 1990*. Am Med Informatics Assoc.

Duvillard EE. 1806. Analyse et tableaux de l'influence de la petite vérole sur la mortalité à chaque age, et de celle qu'un preservative tel que la vaccine peut avoir sur la population et la longevité. Imprimerie Imperiale: Paris.

Furedi A. 1999. The public health implications of the 1995 'pill scare.' *Human Reproductive Update*. 5, 621-626.

Galton F. 1889. Natural Inheritance. London: Macmillan.

Gardner MJ, Altman DG. 1986. Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal*. 286:1489-93.

Gawande A. 2010. Letting Go. Annals of Medicine. *The New Yorker.* 2 August 2010.

Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM & Woloshin S. 2008. Helping Doctors and Patients Make Sense of Health Statistics. *Psychological Science in the Public Interest*. 8:2, 53-96.

Goodman SN. 1999. Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy

Gould SJ. The median isn't the message. *Discover Magazine*. 1985.

Green S, Benedetti J, Crowley J. 2002. *Clinical Trials in Oncology*. Southwest Oncology Group Statistical Center and Fred Hutchinson Cancer Research Center. Seattle, Washington. Chapman & Hall.

Hersh AL, Stefanick ML & Stafford RS. 2004. National use of postmenopausal hormone therapy: annual trends and response to recent evidence. *JAMA*. 291(1):47-53.

Hill AB. 1937. *Principles of Medical Statistics*.

Hoffrage U & Gigerenzer G. 1998. Using natural frequencies to improve diagnostic inferences. *Academic Medicine*. 73, 538-540.

Hooke R. 1983. *How to Tell the Liars from the Statisticians*. Marcell-Deckker.

Hulley S, Grady D, Bush T, Furberg C, Herrington D, Riggs B & Vittinghoff E, PhD. 1998. Randomized Trial of Estrogen Plus Progestin for Secondary Prevention of Coronary Heart Disease in Postmenopausal Women. *JAMA*. 280(7): 605-613.

Jemal A, Ward E & Thun MJ. 2007. Recent trends in breast cancer incidence rates by age and tumor characteristics among U.S. women. Breast Cancer Research.

Looney SW, Grady CS & Steiner RP. An update on biostatistics requirements in U.S. medical schools. *Academic Medicine*. 73:1, 92-94.

Meier P. 1977. The biggest public health experiment ever: the 1954 field trial of the Salk poliomyelitis vaccine. *Statistics: A Guide to the Unknown*. The Joint Committee on the Curriculum in Statistics and Probability of the American Statistical Association and the National council of Teachers of Mathematics. Duxbury Press: Belmont, California.

Neyman J & Pearson E. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philo Trans Roy Soc A*. 231:289-337.

Nguyen A & Fan SK. 2009. Ethics and stopping rules in a phase II clinical trial. *Chance*. 22:4, 39-44.

Sackett DL. 1979. Bias in analytic research. *J Chronic Dis*. 32:51-63.

Schoonjans F, Zalata A, Depuydt CE & Comhaire FH. 1995. MedCalc: a new computer program for medical statistics *Computer Methods and Programs in BioMedicine.* Volume 48, Issue 3, Pages 257-262.

Senn S. 2003. *Dicing with Death: Chance, Risk and Health.* University College London. Cambridge University Press: the Edinburgh Building, Cambridge, United Kingdom.

Sone S, Li F, Yang Z, Honda T, Maruyama Y & Takashima S. 2001. Results of a three-year mass screening programme for lung cancer using mobile lowdose spiral computed tomography scanner. *British Journal of Cancer*. 84, 25-32.

Sterne JAC & Smith GD. 2001. Sifting the evidence: what's wrong with significance tests? *Journal of the Am PT Association*.

Viagra Special Report. 2000. *The Big Issue Magazine*. Vol 350. Melbourne, Australia.

Wakefield AJ, Murch SH, Anthony A, Linnell J, Casson DM, Malik M, Berelowitz M, Dhillon AP, Thomson MA, Harvey P, Valentine A, Davies SE & Walker-Smith JA. 1998. Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, Volume 351, Number 9103.

Wilson RA. *Feminine Forever*. 1966. New York: M. Evans and Co., Inc. Philadelphia: J. B. Lippincott Co. Kathryn S. Huss, MD, Reviewer.

Windish DM, Huot SJ & Green ML. 2007. Medicine residents' understanding of the biostatistics and results in the medical literature. *JAMA*. 298:9, 1010-1017. American Medical Association.