Two Talks by Samuel W. Greenhouse
George Washington University

Abstract

Joel B. Greenhouse
Carnegie Mellon University

The following two papers are written versions of talks given by my father, Samuel W. Greenhouse. The first paper in the series, entitled *On Psychiatry, Epidemiology, and Statistics: A View from the 1950's and 60's,* was presented in 1999 at the Harvard School of Public Health. The second paper, *The Growth and Future of Biostatistics (A View from the 1980's)* was the 1982 invited ENAR Presidential address delivered in San Antonio. It is an honor and a privilege to be able to include them as part of this special issue of *Statistics in Medicine* dedicated to him. Although these talks were not part of the NIH symposium on "Perspectives on the Biostatistical Sciences: A Symposium in Memory of Samuel W. Greenhouse" (June 11, 2001) it seems fitting that in the first NIH biostatistics symposium in which my father did not participate, he is still able to contribute to the proceedings. In the introduction that follows, I provide some background and context for these talks.

Introduction

Joel B. Greenhouse [1]
Carnegie Mellon University

The following two papers are written versions of talks given by my father, Sam Greenhouse, Professor Emeritus, Department of Statistics, George Washington University. It is an honor and a privilege to be able to include them as part of this special issue of *Statistics in Medicine* dedicated to him. Although these talks were not part of the NIH symposium on "Perspectives on the Biostatistical Sciences: A Symposium in Memory of Samuel W. Greenhouse" (June 11, 2001) it seems fitting that in the first NIH biostatistics symposium in which my father did not participate, he is still able to contribute to the proceedings. The purpose of this note is to provide some background and context for these talks.

In late summer 1999, Nan Laird, then Chair of the Harvard University Department of Biostatistics, informed my father that he had been selected for recognition by the Harvard University Institute of Psychiatric Epidemiology and Genetics for his lifetime contributions to psychiatric epidemiology and biostatistics. This was a very meaningful and deeply appreciated recognition. As part of the award ceremony he was invited to give a lecture on a topic of his choice. Instead of a technical talk, he felt the occasion called for something more reflective and personal. He chose to discuss early developments in psychiatric epidemiology and statistics based on his own experiences at the National Institute of Mental Health during the 1950s and 1960s. Shortly after speaking with Nan, my father learned that he had esophageal cancer. Although he would not be able to deliver this talk in person, he nevertheless worked almost daily on a written version which he asked me to deliver for him. [2] The paper that appears here, entitled *On Psychiatry, Epidemiology, and Statistics: A View from the 1950's and 60's*, was almost entirely written by him. One section, on the role of hypothesis testing in mental health clinical trials, is adapted from some of his earlier writings. This paper nicely complements the papers by Butler (2003) and by Katz and Berman (2003) that also appear in this issue.

The second talk that appears here preceded the Harvard talk by nearly 20 years. In 1982, Ted Colton invited my father to present the ENAR Presidential Address. It has been surprising to me the number of people who still remember that talk and encouraged him and later me to publish it. It obviously left a lasting impression on many of those who heard it. The title of that talk was *The Growth and Future of Biostatistics* and I've taken the liberty here of adding a subtitle, *A View from the 1980's*, in part, to help set the context but also for the nice parallelism

[1] This work was supported in part by grants from the NIMH, MH-65430 and MH-30915.

[2] Presented on December 8, 1999 at the Harvard School of Public Health

with the title of the Harvard talk. A central theme in this talk is his concern with the problem of selection bias as it arises in observational studies and in "broken" clinical trials. In effect, he anticipates the growth and development of research in the area of what is now called causal inference.

I have edited these talks lightly to help get them into a more readable format. Specifically, I've added subheadings to help organize the papers and provided bibliographical and biographical notes for reference. The ENAR talk was really in very good shape. I know that Ted was after my father for a number of years to publish this talk. I can only imagine his reluctance to do that was based on his feeling that as a paper it needed "more" work. I hope its appearance in print would not be a source of embarrassment.

For those who are reading my father's work for the first time and for those who knew my father for some part of the 50+ years in which he was an active member of the statistics community, my hope is that you will hear "Sammy's" voice in these papers - modest with respect to his own contributions, supportive of young statisticians, skeptical of new ideas that had not been tested in practice, insightful about fruitful research directions in statistics, and passionate about science, statistics and the NIH. Unfortunately, what is not captured in the written papers is his ever present sense of humor and the extemporaneous jokes that were de rigueur in any Sam Greenhouse talk.

It is with deep gratitude and appreciation that I acknowledge the organizers of the NIH symposium for this significant recognition of my father and his scholarly contributions, and in particular, Ed Gehan who edited this issue and without whose support and commitment these proceedings might not have appeared in print.

References

Butler, R. (2003). Sam's Scholarly Contributions to Human Aging Studies. In press, *Statistics in Medicine*.

Katz, M and Berman N (2003). The Contributions of Sam Greenhouse to Research on Psychiatric Diagnosis and Psychopharmacology. In press, *Statistics in Medicine*.

On Psychiatry, Epidemiology, and Statistics:
A View from the 1950's and 60's

by

Samuel W. Greenhouse

George Washington University Biostatistics Center

1. Introduction and Background

The following comments may appear to be somewhat disconnected and rambling. My intent is to present a perspective based on my experience and my views on the development in the 1950's and 1960's of psychiatric research, psychiatric epidemiology and certain related problems in statistical methods. My career at the National Institutes of Health (NIH) began in 1948 as a mathematical statistician in the National Cancer Institute. I joined a group of five relatively unknown statisticians who had been recruited in 1947 and 1948 by Harold Dorn, a trained sociologist. Dr. Dorn was well known for his epidemiological studies of cancer morbidity and mortality in the United States. However, the technical leader of the group was Jerome Cornfield. Other members of the group who later gained great reputations were Nathan Mantel and Marvin Schneiderman. (Greenhouse 1997)

Cornfield worked on statistical methods and somewhat on theory. Yet his training in mathematics was self-taught. He was an undergraduate history major at Columbia and never had a formal course in calculus. Epidemiologists are well aware of the contributions Cornfield made to epidemiological research and to biomedical research in general. In addition to developing methods such as the use of the odds ratio as an estimate of relative risk in case-control studies, and the development and application of the multiple logistic risk function, the first application of which was to Framingham data, he was also in the forefront of the most important public health problems of the day (see Greenhouse 1982). In the smoking-lung cancer controversy, for example, he was a leader in attacking the arguments that because evidence from case-control studies did not arise from controlled experiments, i.e., randomized studies, little credence was to be placed in the observed relationships being causal.

There is one other area in which Cornfield played a major role, namely, his strong and persistant efforts to get the NIH, particularly the National Heart Institute, to support major multi-center randomized controlled clinical trials. In the short period of ten years from 1948, Cornfield became one of the world's leading biostatisticians. As a result, he was invited to become head of the Biostatistics Department at the School of Public Hygiene at Johns Hopkins University. In this position he succeeded William Cochran who had joined the Statistics Department at this University under Fred Mosteller. (Greenhouse and Greenhouse 1998)

Besides myself[1], the other survivor of the original group of five is Nathan Mantel[2], considered by many to be the real genius. He was creative, intuitive and a superb analyzer of data. Nathan Mantel

[1] Deceased September 29, 2000

[2] Deceased May 25, 2002

approached a set of data with one purpose: what was the investigator asking, and what was the best most efficient way of analyzing the information in the data to provide the answers. Mantel paid little attention to theories or models. His analyses were direct and robust. If methods existed, he used them; othewise he would devise his own. If he felt that the problem did not permit parametric methods, he would use ranks and other non-parametric techniques. (Gail 1997, 1999; Greenhouse 1999)

In those early days, a great camaraderie developed among us. We almost always lunched together. Our luncheons were usually quiet and sociable, as long as we discussed subjects other than statistics. But when we raised statistical topics, almost always there was loud shouting matches without regard to the comfort of those around us. Not all of our discussions at lunch or in the office involved technical statistics. Our most serious, enjoyable and fruitful discussions related to our view of statistics at the NIH, and what should be the mission of mathematical statistics at the NIH.

The mission we were trying to define refers only to the role of mathematical statisticians involved principally in intramural research. One thing was not subject to any debate, namely, we were at the NIH, in accordance with Harold Dorn's directive, in order to provide the best statistical advice in response to questions posed to us by intramural scientists in the laboratory and by investigators in nonlaboratory settings such as epidemiologists. This purpose was a *sine qua non* which we never questioned. A secondary objective was for the mathematical statisticians to conduct research in methodology and theory. Clearly, to the extent to which this research related to our laboratory consultations, either in extending and generalizing existing techniques or in developing new ones, there was no problem. But the interesting question arose as to whether our research should be limited in that way. What if our own research or our readings led to questions *not* immediately related to any laboratory problems? We finally agreed among ourselves, and I believe (but am not certain) that Dorn also agreed with us, that there should be no limitations on the scope of our statistical research.

There is one last point I want to make about our work in the forties and fifties. There was no Clinical Center at the NIH. Almost all of our work was with laboratory scientists: biologists, biochemists physiologists, and the like either in labs with or without animals. In fact, Cornfield worked with Dean Burk in, of all things, photosynthesis. This experience in the laboratory particularly working with scientists doing their research with animals taught us a lot about design issues, namely, factors which could be controlled experimentally in the laboratory but which could not be controlled in a similar manner when working with clinical investigators working with human subjects. I consider this experience to be so valuable that when I spent a sabbatical year here at Harvard in 1981 with the Biostatistics Department I made the suggestion that it might be worthwhile having all doctoral students spend six months in a medical laboratory. It was my impression that this suggestion was not very well received.

2. The National Institute of Mental Health: Organization and People

The major part of my discussion today begins with my career in the National Institute of Mental Health (NIMH). I joined the Biometry Branch of the NIMH in 1954, as head of the Theoretical Statistics and Applied Mathematics Section. The Chief of the Branch was Dr. Morton Kramer, who did his graduate work in Biostatistics under Lowell Reed at Johns Hopkins University. Although in 1946, Congress authorized the creation of the NIMH, the Institute actually opened it's doors on April 1, 1949 with Dr. Robert Felix as its Director. However, the interests and involvement of the federal government in areas of mental health did not start in 1949. Even before World War I, there existed programs in mental health in the inspection of immigrants at Ellis Island and the establishment of the Government Hospital for the Insane in Washington, D.C. which by the way did not become part of the NIMH until 1967. During World War I, Dr. Thomas Salmon an early President of the Americnn Psychiatric Association, served as the Chief Psychiatrist of the American Expeditionary Force. It was due to suggestions made by Dr. Salmon that a Division of Mental Hygiene was established in the Public Health Service. Also it is interesting that a good part of the federal activities in mental health, starting in 1914 and extending to the present, was in the study and treatment of narcotic addiction.

In 1954, only five years after the Institute was established, the NIMH was operating at full steam. It was alive with staff who looked forward to a great future in meeting the mental health needs of the nation as set forward in the Congressional Act of 1946 and signed by President Truman. Bob Felix was a great Director who knew and believed in the major public health and scientific issues facing the Institute, such as developing a training program to produce needed psychiatrists for research, to set in motion procedures for determining the numbers and needs of the mentally ill in the community, and to provide services in the community. Throughout his tenure as Director, Dr. Felix showed a great ability to get things done. This accomplishment was in large measure due to his awareness of the political arenas within which he would have to work: the NIH, the United States Public Health Service, the Department of Health, Education and Welfare, and lastly the various groups that constituted the Institute's clients. Furthermore, there was the support of first rate senior and middle level staff who were highly motivated and believed in the public health and scientific missions of the Institute.

The organization of the NIMH consisted of the Director's office with the usual administrative branches, a major extramural program wherein was the important Grants Management Branch, a number of branches representing various disciplines supported by the NIMH through its grants program, a magnificent training branch responsible for the training and post-graduate education of thousands of psychiatrists, psychologists, epidemiologists, statisticians, social workers, etc., a relatively new Psychopharmacology Branch to develop and sustain a research program to evaluate the therapeutic effects of the recently discovered class of drugs of reserpine and chlorpromazine, and the Biometry Branch which I believe was in the office of the Director. The Intramural Program was a bit unusual for the NIH. The scientific programs of two Institutes, that of the Neurological and the Mental Health Institutes were combined under one Director, namely, Seymour Kety. The NIMH Intramural Program included a Laboratory of Psychiatry, a Laboratory of Psychology, headed by David Shakow who came from the University of Chicago, a Laboratory of Clinical Science, headed by Kety and succeeded by Lou Sokoloff when Kety went to Harvard, and an important Laboratory of Social Psychology headed by John Clausen, and later succeeded by Marion Yarrow. The intramural clinical program was headed by Dr. Robert Cohen.

I note with some regret particularly before this audience that there was no Epidemiology Branch either in the Intramural or in the Extramural Programs. Later, such a branch was established under Deryl Regier and headed mostly by Ben Locke. I assume this came about because Mort Kramer, Chief of the Biometry Branch, was considered to be carrying out the most effective epidemiologic studies appropriate in Schizophrenia and depression for that time (see Ellenberg 1997). Perhaps many of you are not familiar with this research. Kramer with the cooperation of the majority of the State Mental Health Agencies in the country set up a Model Reporting Area responsible for reporting complete statistics on the admissions and discharges of patients in the member State Mental Hospitals. The data so collected were rich in epidemiological information. So for example, Kramer easily began to draw comparisons of duration of hospital stay between married men and single men, between married women and single women, holding various other factors constant. He also was able to obtain conventional incidence and prevalence rates of schizophrenia adusted for age, sex and other factors. Yet throughout the twelve years I was with the NIMH, I used to hear comments about the lack of ”real” epidemiology research at the NIMH. It is not clear what was meant by this phrase. Of course, it was not the kind of risk factor or etiologic research that has become an integral part of current epidemiological research in all chronic diseases. I assume that mental health epidemiology from the seventies on has embarked on the same kind of studies as elsewhere. But it is not clear to me that there has been much progress over and above what Kramer had found from his studies. In fact, Martin Katz who played a leading role in the Psychopharmacology Unit and later as an investigator in the study of the biological aspects of depression has often said there is no marker thus far discovered for schizophrenia or for depression that is equivalent to, say, blood pressure or cholesterol for heart disease. In any event, the Model Reporting Area program was an early example of health services research and became a powerful tool in the hands of Bob Felix when he made his yearly excursions to the Congress to defend his budgets.

One additional word with respect to epidemiologic research. In particular, I must mention a key contribution Dr. Kramer made in psychiatric epidemiology. He was responsible for the development of more reliable diagnostic methods, more systematic and valid reporting schemes for the incidence of mental disorder, and for the construction of centers throughout the country which could conduct epidemiologic research. He was the motivating force behind the now famous joint U.S.-U.K. project on diagnosis, for which he served as Project Officer for the NIMH. By attempting to explain the discrepancies in mental health statistics reported in these two countries, this study resulted in world wide concern for modifying and improving the diagnostic system and helped lay the basis for the development of the empirically based DSM-III in psychiatry. He also participated in the World Health Orgnization International project on Schizophrenia, supported by the NIMH, which resulted in the major revisions of the international classification system (ICD) for mental disorders.

3. Examples of Some Statistical Collaborations at the NIMH

To return to my own activities in the Institute, the section I headed was responsible for providing consulting services to investigators in the intramural program. Seymour Geisser joined the section in 1955 and shortly thereafter Donald Morrison arrived, both from the University of North Carolina. Geisser, by the way, was one of the very few students who received their doctorate with Harold Hotelling. Each of us built up our own group of scientists that we worked with. I consulted a lot with many of the psychologists, social psychologists, and various biochemists and biologists in the laboratory. Some problems I encountered in the intramural program were the development of a model for multiresponse-choice probability learning (Greenhouse, Little, Brackbill, and Kassel 1960), a stochastic process arising in the study of muscular contraction (Greenhouse 1961), models for the interpretation of experiments using tracer compounds (Cornfield, Steinfeld, Greenhouse 1960), multiple comparisons for adjusted means in the Analysis of Covariance (Halperin and Greenhouse 1958), sequential clinical trials with Cornfield (Cornfield and Greenhouse 1967), and profile analysis (Geisser and Greenhouse 1958; Greenhouse and Geisser 1959) which I shall discuss in greater detail next.

3.1 Profile analysis

A common methodological issue which kept cropping up again and again both for Geisser and myself in our consultations with the psychologists was the analysis of vectors of means obtained on different groups. The typical problem was the following: $n(g)$ subjects in the g-th group, $g = 1, 2, \ldots, G$, yield test results on each of k tests or variables. These data summarize into G vectors of means and a pooled covariance matrix, assuming homogeneity of the G covariance matrices. What can be inferred about a common vector of means for the G groups? And if not common, what can be said about the vectors being of the same shape (that is, parallel)? Assuming multinormal theory. a solution was given based on Hotelling's Generalized T-statistic. But a more interesting approximate solution was given utilizing ordinary analysis of variance methods. This latter evidently found great favor with psychologists and others and eventually became a standard part of the output of many popular statistical computing packages. There were two joint papers published: the first in the *Annals of Statistics* (1958) and the second in *Psychometrika* (1959). Twenty years later the latter paper was designated a Citation Classic (Current Contents, July 12, 1982).

3.2 Human aging study

As an illustration of another type of statistical collaboration at the NIMH, I would like to discuss next a major, long-term NIMH multidisciplinary study of ”normal” aging begun in 1955. The goal was to study ”normal” aging, that is, to study elderly individuals living in the community whose health was free from disease. It was the first interdisciplinary project undertaken at the NIH: it brought together psychologists, psychiatrists, physiologists, social psychologists, social workers and of course statisticians. Donald Morrison and I were integral participants from the planning meetings until the end of the project. (Birren et al. 1961; Butler 2002).

The interdisciplinary set of studies were designed to observe results of the ”normal” process of aging in men. It was a longitudinal study where the aims were to identify physical, mental, social/environmental processes that might characterize the disease free elderly patients on hand. Our working hypothesis was based on, ”[T]he optimistic attitude . . . that pathological changes which are occasionally observd in younger individuals and which occur frequently but not uniformly in aged persons reflect the influence of disease rather than that of normal aging. If such extraneous factors could be eliminated then one could examine the changes induced in the organism by aging per se.” (Birren et al. 1961, p. 309)

Volunteers for the study came from two sources: a retirment home in Philadelphia, and from the Association of Retired Civil Employees in Washington D.C. A total of 54 subjects were recruited. Each spent two weeks at the NIH Clinical Center. Initial medical examinations found serious medical disease in seven who were eliminated from the study. The remaining 47 were divided into two groups: Group I, 27 individuals without any observed evidence of disease, or questionable evidence of mininal disease; and Group II, 20 individuals with asymptomatic or sub-clinical disease. There was a host of interesting and apparently important observations and findings just too numerous to mention. I relate only a few. Comparison of medical studies between the two groups and younger cohorts revealed few significant age-specific differences in quantitative variables, except for serum albumin decreased with advancing age; an expected increased blood pressure in the presence of arteriosclerosis; subjects in Group I did not differ significantly in cerebral blood flow and oxygen consumption from subjects fifty-years younger; cerebral glucose consumption was significantly reduced in subjects in both groups; EEG findings showed signifcant differences in both groups from young adults as reflected in a shift to slower activities in the frequency spectrum; one significant correlation tended to confirm previous observations in the literature, namely, a negative relationship between blood presssure and percent slow EEG activity in group II but not in group I (explained by the authors as suggesting that increased blood pressure exerts a protective effect on the electrical activity of the brain when vascular disease is even minimally present).

In summary, I quote: ”When the results of the cerebral circulatory and metabolic and the EEG studies are examined as a whole, they suggest that the brain does undergo change as a consequence of chronological aging per se, more clearly manisfested in its electrical activity than in its circulation and metabolism. However, when arteriosclerosis is present the pathological change in the vascular system becomes the pacemaker of the decline in functions of the brain with age.” (Birren et al. 1961, p. 312)

With respect to psychiatric and psychological evaluations again I quote: ”Compared with the prevailing medical and psychiatric view of the aged, both social psychological and psychiatric interviews revealed these men as a whole to be vigorous, candid, interesting, and deeply involved in everyday living – they were resourceful and optimistic. The group was not uniform, however. Some individuals showed maladaptive patterns of withdrawal and depression. Some showed evidence of mental decline.” (Birren et al. 1961, p. 314)

In addition to these important and impressive results, this study contributed to the development of new subject-matter as well as statistical methodologies and to the practice of collaborative longitudinal research.

4. State of Psychiatry and Psychiatric Thinking

I would now like to make some general comments about the state of psychiatry and psychiatric thinking concerning the evaluation of therapeutic interventions during this period from my perspective at the NIMH.

The psychiatrists in the intramural laboratory were almost exclusively pychoanalysts and psychotherapists. Indeed, psychoanalysis was a favorite topic in those days. This was true to a large extent among the psychologists. This was true despite the fact that the background of the clinical Director, Robert Cohen, was biological and of course Seymour Kety was biological. But there was a certain tension between psychologists and psychiatrists. I recall some of the psychologist wanting to do a project evaluating analysis and psychotherapy within the NIMH but that there was a great reluctance to do so on the part of the analysts. At this point I sympathized with them for given the very small numbers of subjects they had to work with in the Clinical Center, and the long length of time they claimed they needed to obtain results, there was little chance of a successful outcome either way. However, there was an attempt made by another means. An elaborate facility was built to film the sessions between one therapist and one patient. In all, 500 films/sessions were made. Dr. Morris Parloff was involved and he tells me the films were excellent. His description of the project is fascinating (personal communication) and I hope he publishes it somewhere. In addition to the primary task of evaluating the efficacy of the treatment, a major objective was to provide a set of almost real life films for educational and training purposes. A special camera had to be developed in order to film the 50 minute session continuously without having to reload every 15 minutes as was the case with the ordinary Hollywood camera. While the sessions were going on, the group of reviewers were undergoing training with experienced analysts to gain insight in the analysis of films. After some typical group dynamics in trying to arrive at a way of analyzing a session, the group finally settled down ”to work together in a collaborative, constructive, integrative and goal-directed manner.” The shocking thing is that it took them 6 months to complete the analysis of the first 10 sessions. Fortunately for the group the patient terminated the treatment. There are two results worth mentioning. One, as the sessions continued, the analyst became increasingly stressed to produce some signs of improvement in the patient’s condition. To quote Parloff, ”. . . he had begun radically to deviate from standard analytic practice. He appeared not simply to be pressing the ’parameters’ of acceptable techniques but was openly recycling techniques snatched from alien forms of psychotherapy.” The patient was not too clear as to whether there was any significant improvement in her condition. A second interesting result was the criticism raised by some as to the intrinsic value of the data. Again quoting Parloff, ”. . . analysts continued to invoke what they were pleased to refer to as the Heisenberg Uncertainty Principle as invalidating the effort, namely, —the very process of recording and filming psychonanalysis, which involved the full knowledge of the analyst and the patient of the activity had so distorted and destroyed the very subject matter that we could not possibly study true psychoanalysis as it was practiced in real life”.

Interestingly, today it is now standard practice to provide a manual for the delivery of any psychotherapy that not only helps set a standard for clinical practice, but also for research purposes provides a standard for evaluating the proper implementation of the therapy which can be and usually is evaluated using audio or video technology.

Another point of interest, again made to me by Martin Katz, is that the psychiatric state of mind reflected in NIMH activities merely reflected the state of affairs in the outside world. The so called establishment of psychiatry in academia was almost all dominated by analysts and psychotherapists. Starting with the decade of the seventies, a very great change in ideas and practices took place. Departments of Psychiatry in academia became almost exclusively biologically oriented. This shift in emphasis is clearly reflected in the current status of research emphasis at the NIMH where we now have in the intramural program laboratories such as: Biological Psychiatry Branch, Clinical Neurogenetics Branch, Clinical Psychobiology Branch, etc.

5. Statistical Thinking in Psychiatric Research: Clinical Trials

Not until 1956 or 1957 did the Institute established a Psychopharmacology Unit under Jonathon Cole. I served on its Advisory Board from 1957 to 1960. This was a period of intense consideration of the two psychotropic agents: rauwolfia and chlorpromazine. This was also a period of intense activity in planning for clinical trials in two other Institutes: the National Cancer Institute and the National Heart Institute. In fact, from 1955 to 1970, clinicians together with statisticians at the NIH established the basic components, logic and methodology of what later became known as the randomized, controlled clinical trial. The nature of the two different diseases resulted in different procedures: Cancer established groups of investigators according to the class of cancers to be treated; the Heart Institute favored individual principal investigators applying for contracts or grants but all following the same model of clinics, data monitoring and safety boards, policy advisory boards, a data coordinating center and central laboratories. Strangely, despite the existence of the Psychopharmacology Unit, the NIMH never equaled the pace at which the other two Institutes began to support major clinical trials. The NIMH did have an active program of clinical trials under the direction of Jerome Levine. But they were not of the same magnitude. My guess now as to why this happened is that the clinicians and scientists in cancer and heart research readily accepted randomization where as psychiatrists at that time even questioned the need for controls let alone agree to the randomization of patients. As one rather famous psychiatrist said, ”I know this disease (schizophrenia) and I know its course. Why do I need controls?”

In addition, I encountered at the time a general skepticism among psychotherapists for the need to develop or refine methodology for doing research in the evaluation of therapies in mental disorders. Thus, for example, one would encounter attitudes such as the following :

> I do not think that problems in methodology are usually of prime importance in doing research. The refinement of methodology, while a good thing in and of itself, should be secondary to the discovery of relationships and the development of understanding. (Festinger, 1959)

I suspect that attitudes such as this and the fact that in these early years psychiatric and psychopharmacologic investigators turned to psychologists to provide statistical and methodological support for clinical trials contributed to the delay in the advancement of methods in these areas.

5.1 The role of hypothesis testing in clinical trials[3]

In the late 1950's and early 1960's, some of the senior statisticians at the NIH began to consider issues related to the design and conduct of clinical trials. The National Cancer Institute, for example, had already been involved in carrying out small clinical trials in its chemotherapy program and the Heart Institute was beginning to plan large multi-center trials. We were also aware of those clinical trials being conducted in England under the auspices of the Medical Research Council. Initially, based on our early experiences with collaborations with laboratory scientist doing animal studies, we thought of the clinical trial as a mere extension of the principles of good research design and analysis. It was not long, however, before we realized that clinical trials were complex and difficult modes of research in all areas of design, implementation and anaylsis. To further explore these issues of what is now called "statistical practice" (i.e., where statistical theory and methods interface with real applications), the statisticians at the NIH in 1965 held an informal seminar to discuss the role of hypothesis testing in clinical trials. Cornfield spoke about his experience with trials in cardiovascular disease, Marvin Schneiderman in cancer trials, and I spoke about trials in psychiatry. Marvin Zelen was the formal discussant. The following is an excerpt from my comments which, I believe, still reflect on issues of implementation of clinical trials in psychiatry today (Cutler, Greenhouse, Cornfield, Schneiderman 1966).

> In research of therapies for the mental disorders, it is very rare that an investigator enters into a clinical trial with only one question to be answered. In this area competing drugs tend not to cure diseases but lead to improvement in disease status. They alleviate symptoms. Some drugs are tested for their efficacy in reducing hostility and hyperactive behavior. Other compounds are tested for their efficacy in removing depressive symptoms. Within any set of comparable drugs, many of the compounds show some effective properties. Hence, when a new drug is put forth, the psychiatrist or psychopharmacologist wants to know much more than the answer to the question whether it produces significantly more improvements in patients than standard drugs and how much more. There are other major questions on which the trial must provide information: are the symptoms affected by the new drug the same as those alleviated by the known drugs; what are the effects on the different variables that are being observed simultaneously and do interactions exist among these effects; how do the effects vary with time on treatment.
>
> The existence of multiple major questions in a clinical trial is exemplified in a recent

[3] NOTE: Although Sam did not write the following for this particular talk, it is taken from his other writtings and reflects the points and issues he wanted to discuss here. JBG

study, involving the collaboration of nine hospitals, carried out by the NIMH-Psychopharmacology Service Center (1964). The drugs being tested were two new phenothiazine compounds and the older, relatively well known drug, chlorpromazine. I quote directly from this report to list the questions constituting the major objectives of the trial.

1. What proportion of acute schizophrenic patients show clinically significant improvement on phenothiazine treatment? Even after improvement, to what extent are patients still mentally ill?
2. Do the active drugs differ in their effect on specific schizophrenic symptoms? For example, is chlorpromazine more effective in reducing hostility, and fluphenazine more effective in reducing withdrawal?
3. Are two newer phenothiazines, thioridazine (Mellaril) and fluphenazine (Prolixin) more effective than placebo, and are they as effective as the older standard phenothiazine, chlorpromazine (Thorazine), in the treatment of acute schizophrenic patients?
4. Are there differences between the drugs in the nature and/or frequency of the side effects produced?

In psychotherapy, one rarely speaks of cures but rather of improvement. Is there one agreed-upon measure, by which improvement is assessed? In 1956, shortly after the discovery of the behavior-damping effects of certain phenothiazine and rauwolfia compounds, I made a study of 25 or so controlled studies. In 11 trials evaluating only chlorpromazine on psychotics, I found the following criteria measurements used: the Ferguson scale, the Lorr scale, the Malamud-Sands test, the Wechsler-Bellevue Intelligence scale, the Fergus Falls Behavior Rating Scale, the Minnesota Multiphasic Inventory, and a good old-fashioned clinical assessment of disease status by a psychiatrist, although it never is clear how dependent or independent this clinical rating is of the objective tests measured. It was not unusual in a study to find mutliple assessments of outcomes but also mutliple assessors. For example, in one study, symptom and behavior assessments were rated by the doctor on the basis of interviews and by the nurses on the basis of observing the patient on the ward.

How does one fit these mulitple objectives and measurements of response into the traditional hypothesis testing framework? Does one set the same type I and type II error levels for each test separately? Which measurement determines the choice of sample size? In the actual significance testing how does one draw a conclusion? Do you get a combined, weighted P-level? What if one observes anomalies such as the fact that the treatment produces a significant improvement in one measurement but not in another? The literature has many instances where improvement was noted on the basis of clinical judgment but no improvement was observed on objective tests. Is this merely due to bias of the clinician? The answer is not clear. The different tests and the clinician may be tapping different facets of mood, behavior, organic defects etc. and the treatment may really vary in its effect on each of these.

But in research on therapies of mental disorders it is just this information which may turn out to be most important in clinical trials. It could lead to greater insight of the measurement instruments and to a finer classification of schizophrenic patients previously believed to suffer from the same disease. In fact, the broader approach to clinical trials in psychopharmacology led to increased research in diagnosis and prognosis (Katz 1966; Zubin 1987).

Another aspect of clinical trials in psychopharmacology is worth mentioning. Historically, the initial enthusiasm for chlorpromazine stemmed from uncontrolled studies, what we tend to call ’clinical impressions.” The more scientifically minded psychiatrists began demanding controlled studies. This lead to trials which met sound statistical requirements: random assignment of patients to treatment and placebo groups and double-blind testing, where neither the subjects nor the clinician doing the rating knew who was on drug and who was on placebo. It was not long before criticisms of the ‘blind’ aspects of these studies began to appear. Evidently chlorpromazine, reserpine, and other agents produce minor side-effects within an hour of administration. The most important of these side-effects was drowsiness. Thus it was easy in many cases to discover those who were on the drug. The reaction to this phenomenon was to employ placebos which produced similar side-effects. Furthermore, these early studies led to another interesting finding. Relatively large numbers of patients on placebo showed signs of improvement ranging from fair to considerable. In the survey referred to earlier, for example, the improvement rate among psychotic controls varied from 17 to 36 per cent.

In presenting this discussion, I have intended to demonstrate that most clinical trials are extremely complex phenomena. It therefore seems to me highly undesirable to make the sole conclusion to be drawn from a clinical trials the rejection or acceptance of the null hypothesis. The investigator must be allowed the flexibility of drawing additional appropriate conclusions based on information provided by the data even though the issues where not hypothesized before the trial.

6. Concluding Remarks

A few last remarks with regard to the “Statistics” of the title.

Everitt in a review of the use of statistics in psychiatry in the mid 1980s made the point that psychiatric researchers would benefit immensely from applying some of the techniques used in other fields such as logistic regression, survival analysis, log-linear models and the like (Everitt 1988). I was a discussant on that paper. I very much agree with him on this main point. But I still believe improvements can be made in measuring instruments. I made the point that before seeking new techniques for data analysis psychiatrist and psychologists should reconsider some basic issues in the measurements they use: for example, seven point scales used in correlations, etc. It is this very important issue which to me answers the question that I raised, namely, “In what sense does statistics in psychiatric research differ from statistics in other fields of application.” Evidently this problem of measurement is the first thing that strikes statisticians with experience in data analysis in cardiovascular, cancer or diabetes research. For example, I have heard many refer to data from mental health studies as soft.

6.1 The future

The area of greatest advance in psychiatry that will occur in the near future, I am convinced, will be related to genetics. And it may very well be the case that it is psychiatric genetics, and in particular, psychiatric epidemiology that will be the driving force for either new methodological techniques, or more imaginative uses of standard procedures using more complex models. The problems identified with the mental disorders on the molecular level pose considerable challenges for effective analysis. The recent supplement to the *Lancet* (July 1999) contains an excellent article on psychiatric genetics by M.J. Owen and A.G. Cardno. The authors succinctly summarize all the issues: the action of multiple genes each of which contributes a small amount of change as well as environmental effects; the possibility of non-additive effects such as gene-gene interactions and environment-gene interactions; the particular problem in psychiatry where genetic complexity is enhanced by the complexity of phenotype; the use of statistical methodologies used in previous work on single gene disorders. In terms of familial studies, the authors point out the attempt to map genes of small effect, “the number of families required for linkage studies becomes prohibitively large.” Clearly, the need to refine phenotypes, to make more precise diagnoses, to identify modifying genes and most importantly “the implications of identifying genetic risk factors for the major psychiatric disorders” will require a complete review of the kinds of analytic techniques to bring to bear on decisions regarding designs of studies and optimum methods in data analysis. In my view, it is this area in which much of new statistical thinking will develop.

References

Birren, J., Butler, R., Greenhouse, S.W., Sokoloff, L. and Yarrow, M. (editors) (1961). *Human Aging: A Biological and Behavioral Study*. Public Health Service Publication No. 986.

Butler, Robert (2003). Sam’s scholarly contributions to human aging studies. *Statistics in Medicine*, xxx:sss-sss.

Cornfield, J. and Greenhouse, S.W. (1967). On certain aspects of sequential clinical trials, in *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 5:813-829.

Cornfield J, Steinfeld J and Greenhouse S.W. (1960). Models for the interpretationof experimentsusing tracercompounds. *Biometrics*, 16:212-234.

Current Contents (July 12, 1982). This Week’s Citation Classic, 14:20.

Cutler S, Greenhouse S.W., Cornfield J, Schneiderman M.A. (1966). The role of hypothesis testing in clinical trials, *Journal of Chronic Disease*, 19:857-882.

Ellenberg, J., (1997). A conversations with Morton Kramer. *Statistical Science*, 12:103-107.

Everitt, B.S. (1987). Statistics in Psychiatry (with discussion), *Statistical Science*, 2:107-134.

Festinger L. (1959). Sampling and related problems in resarch methodology: Proceedings of the 1959 Woods School conference on approaches to research in mental retardation. *American J. of Mental Deficiency*, 64:2.

Gail, M. H., (1997). A conversations with Nathan Mantel. *Statistical Science*, 12:88-97.

Gail, M. H. (1999). Some of Nathan Mantel's contributions to Epidemiology. *Statistics in Medicine*, 18, 3389-3400.

Geisser S, and Greenhouse S.W. (1958). An extension of Box's results on the F distribution in multivariate analysis. *Annals of Mathematical Statistics*, 29:885-891.

Greenhouse, S.W (1961). A stochastic process arising in the study of muscularcontraction, in *Fourth Berkeley Symposiumon Mathematical Statistics and Probability*, University of California Press, 4: 257-265.

Greenhouse, S.W. (1982). Jerome Cornfield's contributions to epidemiology. *Biometrics Supplement*, 38, 34-45.

Greenhouse, S.W. (1997). Some reflections on the beginnings and developments of statistics in "Your Father's NIH, *Statistical Science*, 12:83-87.

Greenhouse, S.W. (1999). A selection of Mantel's contribution to laboratory research. *Statistics in Medicine*, 18, 3401-3408.

Greenhouse S.W., Little K.B., Brackbill Y and Kassel S.H. (1960). A model for multiresponse-choice probability learning, (abstract). *American Psychology*, 15:456.

Greenhouse S.W. and Geisser S. (1959). On methods in the analysis of profile data. *Psychometrika*, 24:95-112.

Greenhouse, S.W. and Greenhouse, J. "Jerome Cornfield." In: *The Encyclopedia of Biostatistics*, eds. Armitage, P. and Colton, T. John Wiley, New York, 1998.

Halperin M, and Greenhouse S.W. (1958). Note on multiple comparisons for adjusted means in the analysis of covariance. *Biometrika*, 45: 256-259.

Katz M (1966). A typological approach to the problem of predicting response to treatment. In *Prediction of Individual Differences in Response to Pharmacotherapy*, J. R. Wittenborn and T.R. May (editors). C.C. Thomas, Springfield, Il.

NIMH-PSC Collaborative Study Group (1964). Phenothiazine treatment in acute schizophrenia. *Archives of General Psychiatry*, 10:246.

Owen MJ, Cardno AG. (1999) Psychiatric genetics: progress, problems, and potential. *Lancet*. Jul;354 Suppl 1:SI11-4.

The Growth and Future of Biostatistics

(A View from the 1980's)
by

S. W. Greenhouse
Harvard School of Public Health

March 24, 1982

1. Introduction

When Dr. Colton initially invited me to give this address, I said no. In fact I said no over a period of two weeks. In the course of our conversations, it became clear that he wanted me to expand on some-comments I had made in a recently published paper dealing with epidemiological issues for the 80's (Greenhouse 1980). These remarks related to observational studies. Specifically, I indicated that for prospective observational studies where two treatments are being compared in the absence of randomization the most serious defect in drawing valid inferences were the unknown effects of selection bias – physician selection and patient selection. I noted that although a good deal of research had appeared due to Cochran, Rubin (e.g., 1973), and others in trying to adjust for bias due to imbalances in baseline covariates, little or no research had been done on the potential selection bias. I pointed out that in the area of sample surveys there seemed to me to be a potential bias of a similar nature, due to non-response and yet a fair amount of research on possible effects of this cause of bias had been engaged in by survey statisticians. It was this topic that Ted wanted me to expand on. I told him that unless I had something definite to offer in ways of conducting this research there was no point to repeating these comments on the need for such research. He then twisted my arm to accept, think about the problem, and also have the option of making any other remarks I thought appropriate. Thus the general nature of the title. In the next thirty or forty minutes I shall make some personal observations on biostatistics and present some thoughts on observational studies.

By reflecting on biostatistics, I mean to present my impressions on several aspects of the discipline called biostatistics. These refer briefly, certainly not in depth, to the recent historical development of the subject, to the relatively recent association with epidemiology, to the relationship between biostatistical problems and theoretical statistics and lastly to some views on training. These contemplations must of course be based in my own experience beginning with the year 1948, when I first went to the National Institutes of Health (NIH), up until the present, a period of 34 years.

2. From Biometry to Biostatistics

Dr. Harold Dorn, a trained sociologist conducting epidemiological research in cancer at the NIH, recruited five statisticians in the years 1947-48, Cornfield, Mantel, Lieberman, Schneiderman, and myself. Max Halperin was brought on in 1951 as the mathematical statistician in the recently established Biometrics Research Branch of the NIH. This group of six statisticians has at times been referred to as the original group. Dr. Dorn and the five of us in 1948 were part of the U.S. Public Health Service and housed on the NIH grounds. Within a year or two, the group was transferred to the newly created National Cancer Institute (NCI) and formally established a Biometry Branch with Dorn as the chief. As the different categorical institutes were formed, almost all provided for statistical research and services in Biometry or Biometrics Branches. The question of interest is why Biometry? Why not Biostatistics? The answer for the institutes following the NCI is obvious. It was simply easier for organizational approval and uniformity to follow the structure set by NCI. The factors motivating Dorn, Cornfield and the rest of us to call the branch Biometry is somewhat more complicated. Incidentally, I depend upon a rather poor memory but I do not believe the alternative to Biometry was ever Biostatistics. In fact I am not sure there ever was an alternative. Professionally, there existed the Biometrics Section of the ASA which began publishing the *Biometrics Bulletin* in 1945. The International Biometric Society was formed in 1948-49 and took over the publication of the *Biometrics Bulletin* from the ASA in 1949-1950 which became the journal *Biometrics*. We were very much aware of the fact that the most important journal for publishing papers on applications to the biological sciences was *Biometrika*. But more significant was the fact that all of us had already become deeply involved in consulting with and collaborating with laboratory investigators in the two years before our transfer to the NCI. The range of scientists involved was quite broad – physicists, chemists, biochemists, physiologists, pharmacologists and biologists. At that point we really believed that the primary workscope for us would be what was enunciated in the declaration of the Biometric Society as an international society “devoted to the mathematical and statistical aspects of biology.” When I moved over to the Biometry Branch of the National Institute of Mental Health to head a statistics section, I recruited a biomathematician who was trained under Nicholas Rashevsky (1899-1984), a pioneer of mathematical biology at the University of Chicago. So clearly in those early years we intended to engage in biometry and we actually did so.

This was the situation until the NIH Clinical Center was built and all the Institutes began to develop extensive clinical programs. This fact plus the ever increasing requests for consultations with clinical investigators outside of the NIH, began to change the nature of our interests. Even before the Clinical Center, some of us had become involved in methodological issues in epidemiology. After the rise of the clinical programs, most of our efforts were directed to clinical research and less to laboratory research. For some reason not clear to me these developments somehow fell more naturally under the title of biostatistics rather than biometry.

What was going on outside of the NIH? In academic institutions in 1947-48, there were three well known Departments of Biostatistics in Schools of Public Health (or Hygiene) at Johns Hopkins, Harvard, and Columbia University. I believe, although I may be wrong, that the Departments of Biostatistics at North Carolina and at Berkeley came after 1948. There was a Department of Biometry at Case Western Reserve. In non-academic medical research institutions, I believe the only one that had a statistical unit was the Mayo Clinic. In England there seemed to be no facility titled “biostatistics.” It was either biometry or simply statistics.

The point of all this is that aside from differences in name there was essentially no difference in the research content conducted under the heading of biometry or under biostatistics. In the Schools of Public Health or the Medical Research Institutes there may have been more emphasis on medical problems, and in non-medical institutions such as the Statistical Laboratory at Iowa State there was a greater concentration on agronomy, but still, statisticians in either facility could just as easily work on the same problems under the heading of biological sciences. Since the late fifties and early sixties, however, there is a definite divergence in the fields of interest between biostatistics and biometry. The former definitely and clearly became almost exclusively involved in epidemiologic and clinical research. Biostatistics took off in its own directions and instead of being either equivalent to or a part of biometry, became an independent discipline or science. The difference in research interests between biostatistics today and, say 30 years ago, is reflected to some extent, although not very markedly, between the type of papers published in *Biometrics* say in 1951 and 1981. From Table 1 it is clear that there has been a decrease in papers on formal designs and the analysis of variance and not just an increase but an addition to the literature of papers on clinical trials, survival analysis and medico-clinical techniques.

Today there are many more departments of biostatistics in schools of public health than existed in the forties. A number of statistics departments, obviously at universities that do not have schools of public health, provide a concentration in biostatistics related courses and seminars for their graduate students interested in biomedical applications. An example of such a program is the one at Stanford. Of unusual significance is that some of the most exciting theoretical work in the design of clinical trials and the analysis of survival data being produced in the United States is taking place in biostatistics departments in schools of public health, and I note in particular, the departments at Harvard and the University of Washington. At the former where I have been taking a sabbatical this academic year, I found what I consider to be the greatest cluster of young (under age forty) statisticians in the world. As a group, they have already contributed greatly to the design and analysis of problems of clinical trials and to other areas in medical and epidemiological research. Similarly, on the same level of high quality theoretical work, Breslow and Prentice have consistently made important contributions to survival analysis and to epidemiological methodology. There are other Biostatistics Departments where individual statisticians have made and are making contributions on a theoretical level. I emphasize here theoretical contributions because in the past one would not ordinarily have expected such work to come out of a biostatistics department. It is also of interest from the point of view of training to point out that almost all of the statisticians I have in mind were trained as theoretical or mathematical statisticians.

3. Biostatistics, Epidemiology and Schools of Public Health

The foregoing remarks have a purpose, namely to establish the fact that biostatistics has evolved into an independent discipline, no longer a part of biometry.

The interesting question now is whether this is a real, solid development with a potential for a long future as an academic discipline or whether it is merely a temporary phenomenon. Part of the answer lies in two directions: one, the relationship between biostatistics and epidemiology and the other the nature of schools of public health.

Table 1: Major Topics appearing in *Biometrics* in 1951 vs. 1981

	1951	1981
Design & ANOVA	10	2
Bioassay	2	4
Genetics	1	4
Diagnostic Tests	2	2
Applications and Agronomy Biol.	3	10
Accident Proneness	1	–
General Methodology	4	10
Clinical Trials and Survival Analysis	–	11
Bayesian and Decision Analysis	–	3
Tracking	–	3
Applications-Clinical	–	6
Epidemiology	–	3
Miscellaneous	4	–
	27	58

The connection between biostatistics and epidemiology has always been close. It will be recalled that up until recently epidemiologists were physicians who favored the specialty of investigating the way diseases occurred in populations, their causes and their relationship to medical and non-medical factors. The problems they attacked were not confined only to the study of epidemics but extended to the evaluation of therapies. Many were skilled in quantitative reasoning and were knowledgeable in the statistical methods of the day. Then beginning with the thirties, epidemiology turned its attention to the study of chronic diseases. It became impracticable to use the same prospective research strategies that were so obviously appropriate in the study of infectious diseases. Imagine a study to determine whether a given agent or set of agents causes a cancer in a specific site where the yearly incidence is say 5 per 10,000 population. One would require a group of approximately 200,000 individuals who are not exposed to the agent in order to observe 100 cases. Since most cancers have a long latency, 15 to 20 years, and assuming the onset of clinical disease occurs with a reasonable frequency at age 50, one would have to accumulate cohorts of 400,000 30 or 35 year old individuals and follow them for 15 or 20 years. It is no wonder that epidemiologists turned to the method of sampling on the occurrence of the disease, that is, the case-control study, to study the chronic diseases. But here it was statisticians, primarily Cornfield and Mantel, who provided a rationale for the approximately valid inference based on case-control data. Biostatisticians then became heavily involved in elaborating on the conditions for valid inferences with concerns about bias due to possible confounding factors. Furthermore, biostatisticians began exploring other issues related to epidemiologic research such as models for determining the effects of possible risk factors for disease.

At the same time another development occurred in epidemiological training which is relevant to the interface between biostatistics and epidemiology. When training grants in general and to epidemiology departments in particular began to decrease and almost disappear, fewer and fewer physicians applied to epidemiology departments for the MPH and DrPH degrees. I believe this trend began about 1965. It was then that these departments began recruiting non-M.D. students into their graduate masters and doctoral programs. Most of these students obtain a fairly good training in biostatistics. A consequence of this development has been an overlap in the research work conducted by graduates of epidemiology and biostatistics programs. There are a number of statistical positions in the cancer centers, in clinical trials coordinating centers and in other programs that can be filled by either group of graduates. Hence, it is possible that the future of biostatistics which clearly must depend upon the numbers of students entering the graduate programs may be threatened by this competition for jobs.

However, there is another more important matter which might jeopardize the continued vitality of biostatistics. If we assume that this discipline is inextricably woven with the strength and welfare of the departments of biostatistics then we are faced with a bizarre academic situation. As far as I know, no school of public health is fully supported financially by its university. ”Not fully” is an overstatement. Actually this fiscal support is almost nil. The extent varies from university to university. Hence biostatistics departments as well as all other departments in the school must look to the outside – so called ”soft money” – for support of their activities. Almost all of this support has in the past come from the federal government through the NIH. Some departments have had as much as 80 to 90% of their budgets come from the NIH. As everyone knows, this NIH support is decreasing. No one knows what the lower limits of this support may be. In schools of public health associated with state supported universities, funds appropriated by state legislatures have also decreased. What the future of financial support will be is anybody’s guess. But clearly biostatistics will be very much dependent on that future. The threat is both to faculty most of whom are non-tenured and to students.

The need for soft money is a direct consequence of the status of the schools of public health in the university. One can only reason that decades ago the public health schools were very small and fully supported by the administration. However, in the fifties and sixties, a period of greatly increased support of medical and public health research by the NIH, faculties expanded with the blessing of the university administration since this growth cost them relatively little. But in this same interval some of the most serious public health problems were developing in the nation – environmental pollutants, low dose ionizing radiation, food additives, occupational hazards, etc., the very problems which the epidemiology, biostatistics, environmental health and other departments in the public health schools were best able to attack. From this point of view I would claim schools of public health have outgrown whatever tradition was responsible for their step-child status. It is time that they become an integral part of the university.

4. Biostatistics and Theoretical Statistics

To return to our main theme of biostatistics as a discipline, I do want to comment on a development which was forced on statistical theory prevailing at the time. The first large scale multi-center, randomized, controlled trial sponsored by the Heart Institute, the Coronary Drug Project, began in the early 1960s. By this time, many of the statisticians at the NIH had already become familiar with the problems in design and analysis of clinical trials. None of us, except possibly Ed Gehan, was personally involved in the operations of any trials. Rather, we were close to their planning and implementation in advisory capacities. Cornfield worked with an early trial comparing two anticoagulants in the treatment of acute myocardial infarction, another testing the efficacy of estrogen in the secondary prevention of coronary disease and others. Gehan was the statistician on a sequential trial in acute leukemia. I was consulting on a trial comparing two new phenothiazine compounds with chlorpromazine in the treatment of schizophrenia. Thus, the topic of clinical trials was a ”hot” issue at the NIH in those days and all the statisticians were interested in the many methodological as well as practical problems arising in those trials. One of these had to do with an agreed upon need to look at the data several times before a trial ended. That is, there was agreement among us at the NIH that it was necessary to monitor the data and it did not need too much persuasion on the part of clinicians to convince us. Clearly, if a new agent was exhibiting severe adverse effects or if a therapy was demonstrating an unusual superiority, some action should be taken. But this kind of behavior violated some basic principles in the classical statistical theory and practice then prevailing. For one, the predetermined Type I and Type II errors are violated. For another, how can we use the very data suggesting a hypothesis to test that hypothesis? Of course much research has been done on these matters since – adjusted p-values, sequential repeated testing rules, likelihood principle inference and Bayesian inference.

In 1965, a symposium on the topic of hypothesis testing in clinical trials was held at the NIH (Cutler et al. 1966). One of the speakers [Greenhouse] ended his presentation with these comments (p. 861-862): “What then is the role of hypothesis testing in clinical trials? I would say the classical precepts of the specification of the two types of possible error and their relationship to the determination of sample size should serve as a guide to help make decisions in the planning stage of the study. As such, the framework can be most useful. But it should not bind the investigator or the statistician in the analysis of the data nor in the information obtainable from the data.”

Clearly such heretical thoughts could not possibly be acceptable to the purist mathematical statistician in the Neyman-Pearson framework. Yet I do not recall hearing any public pronouncements against the methods being practiced in clinical trials since 1960. Cornfield was writing a series of papers that attempted to set a theoretical justification for these practices based on the likelihood principle which was independent of the stopping rule (Cornfield 1966a, 1966b, 1969). But the practices biostatisticians were engaged in did not seem to elicit any response from theoretical frequentists, until very recently.

I refer to a technical monograph by David Hinkley from the University of Minnesota dated January 1982 (see Hinkley 1983). After emphasizing that the mathematization of statistics was very beneficial, he goes on to say,

> ”And yet the formal structures of mathematical statistics seem to have a blind spot. Most

> of the mathematical development has to do with pre-data analysis: Is such and such likely to be a good procedure? How should we plan to do so-and-so? To answer such questions requires one to place one's particular statistical problem a priori in a sample space with superimposed probability distribution over potential realizations. The blind spot is the implicit assumption that pre-data probability calculations are relevant to post-data inferences. This blind spot is covered by Bayes' theorem, which explicitly recognizes the difference between pre-data and post-data contexts. Must the blind spot remain in frequentist statistics? . . . Once exact models are established, theoretical discussion focuses on exactness: unbiasedness, sufficiency, locally most powerful, inadmissibility, ancillarity and so on. Of course these exact concepts are useful, in part because they prevent loose thought and ad hocery – but the concepts and exact properties should be used only as guides for careful approximate analysis."

Note how close this is in thought to the quotation above at the NIH symposium, even using the same term ”guide”. Here, finally, are frequentist theorists coming to grips with the biostatistical practices of the past 20-25 years. Now I have no direct evidence that it was clinical trials that led Hinkley to this awareness of the inadequacy of classical theory for analyzing experimental data. What we do know is that statisticians working in biostatistics were aware of this and it's a good bet that Hinkley was familiar with their problems.

Thus the point has been made that biostatistics has brought about the conditions for a reassessment of classical Neyman-Pearson theory.

5. Biostatistics Training

It may sound strange to many and no doubt there are many who will disagree with me but I think the subject of training of biostatisticians is a different [sic] problem. In order to transmit some of my views let me rephrase a previous phrase from ”training of biostatisticians to training students to work in biostatistics.” The unquestioned fact is that the major advances in biostatistical methodologies have been made by statisticians trained in mathematical statistics departments. This does not mean that biostatisticians in biostatistics departments will not make important, innovative contributions in the future. But it does raise basic questions as to what is the purpose of graduate education in a biostatistics department. I have a feeling every department chairman would like to turn out statisticians well trained theoretically. But to be able to do so one must have students with appropriate mathematical backgrounds and then one wonders why wouldn't such students go to a theoretical statistics department. Indeed one may even go one step further and argue that perhaps biostatistics departments should make their primary function research and confine their teaching activities to provide suitable service courses for other departments within the school of public health. Of course if the argument is made that the objective in graduate training in biostatistics departments is to turn out competent practitioners –statisticians who will be able to design research plans and be able to apply appropriate analytic techniques – this is another matter. This of course is the traditional view of training in the biostatistics department. Then from the point of view of future growth one must take into account the possible limiting factor of the number of positions that can become available in biomedical research.

But philosophy aside, there are two rather minor observations I have about current training in graduate programs in biostatistics. The generality of these observations may be very limited since I am not familiar with the details of training in all departments.

The first pertains to an absence of a course in sampling on the level of Cochran's text (Cochran 1963). I would make this a required course. I do not see how students can really appreciate the nature of case-control research or the various issues relating to observational studies without such a course. The second relates to the opportunity for students to get actively involved in consulting on laboratory research. Perhaps this can be done during one or two summers or during one academic year. There certainly should be facilities available in the laboratories of the nearby medical school. It is not that simple to specify the benefits to be derived from such an experience. They may not be the same for all students nor for all laboratories. But I do have the definite feeling that all of us at the NIH in those early days gained immeasurably from our close association with laboratory research. For one, the student gets acquainted with the "real world" more easily in the sense that many statistical models are more realistic as a result of greater control over confounding conditions. For another, spending even part time during an academic year in two different laboratories, say, a pharmacology lab and a physiology lab, a student will be faced with a variety of statistical problems and the nature of statistical data that can only lead him or her to appreciate that part of statistics which is more an art than a science – to be able to get at the information in the data needed to answer a question, to summarize it, and to find optimal methods of analysis.

6. Observational Studies

I shall now turn to observational studies. It is difficult to argue for something positive with regard to these studies without being accused of being non-scientific, turning the clock back, etc. So let me say at the outset that I will never suggest a non-randomized clinical trial for evaluating therapies as an alternative to a randomized clinical trial (RCT) if the latter can be conducted. On the other hand, may I note parenthetically, that I doubt if I would ever recommend a randomized trial to ascertain whether a suspect factor is etiologic for a disease. But yet, there may be a clinical problem where an RCT is almost impossible. For example, recently papers have been published providing evidence that slightly overweight individuals are healthier and fare better in combating chronic disease than slightly thin or definitely-thin individuals. The benefit claims are quite general affecting a number of health properties. Suppose this finding was to be subjected to a formal study. Would random allocation be feasible? How would we randomize and to what groups? We might decide on three groups among a cohort of say 30 year old males, the groups being as follows: 1) attaining a weight 10% above normal; 2) attaining normal weight and 3) attaining a weight 10% below normal and adhering to one's weight group for a given duration, 10 to 15 years, say. Will we have to stratify by current weight and then randomly allocate to the three groups within strata? This may be important because maybe the effects observed hold only for those who have been in these weight groups since childhood. How do we maintain compliance to treatment groups? How do we assure the single blind, namely for the examining physician who need not be told what groups patients were assigned to? It is not impossible to carry out a randomized trial but the difficulties are such that one must carefully weigh gains and losses between a randomized and non-randomized study. It is important to recognize that not all clinical questions are as clear cut as whether a new therapy is more effective than a standard. I should mention that about four weeks ago my colleagues at the Biostatistics Department at the Harvard School of Public Health brought to my attention a very nice book on observational studies by Anderson, Auqier, Havek, Oakes, Vandaele and Weisberg (1980), titled *Statistical Methods for Comparative Studies*. These authors attempt a classification of comparative studies and give four major reasons for the use of non-randomized studies: 1) ethical, 2) only kind possible, 3) less expensive, and 4) closer to real life situations.

In my opening remarks I referred to the possibility of doing research on selection bias in observational studies. Let me now try to specify the kinds of inquiries I have in mind. I do not include case-control studies. These have been given ample excellent attention with regard to their logic, methodology and inferential characteristics. There are other types. The occupational hazard problem where it is desired to find out whether a group of workers subjected to high levels of some agent are at high risk for a disease. The plaguing issue of the health effects of environmental pollutants where a number of geographic regions –cities, counties, etc. – are selected on the basis of as wide a range as possible of air pollutants and then the inhabitants are followed and observed for various disease outcomes. The former can be longitudinal or retrospective in the data collection sense. The latter is almost always longitudinal.

However, what I want to focus on is the observational study of determining the efficacy of a treatment or comparative efficacy of two treatments. The very kind of study I indicated earlier should be done with randomization and controls. I do this for several reasons. One, I am familiar with one major medical problem concerning treatments where it is urgent that a comparative study be done but where it appears clinicians will not, or let us say, are very reluctant to randomize to a control therapy approved by the FDA. Second, the observational study in medicine is very old, has a most interesting history, and has led to many mistakes but has also produced some great successes.

Early History

It may surprise many to learn that the first clinical trial was not done in the 20th century or that Sir Bradford Hill's book in 1937 was not the first to prescribe in writing the principles to which we all adhere today for the scientifically appropriate evaluation of therapies. What Bradford Hill (1937) added that was not a part of the practice or thinking of the past was of course the element of random allocation. As has been said medicine and clinical observation on disease and the effects of therapies go back to ancient times. One can speculate that there may have been practitioners of the art of medicine from ancient to modern times who were able to reason the principles of sound evaluation of treatment, even the need for randomization, however vague that concept may have appeared to them. To reject this possibility with a wave of the hand is to underestimate terribly the reasoning capability of the human mind from the Babylonian-Egyptian civilization to modern times.

Modern times for our subject seems to have begun with John Lind (1753) some 240 years ago. The story is well known. Lind, a Scottish naval surgeon, decided to do a comparative trial of the then current "cures" for scurvy. He took 12 cases of scurvy out to sea. I quote, "their cases were as similar as I could have them." They were all put in the same area devoted to the sick and all were given a common diet. He divided the patients into six groups of two each. Each of five pairs was treated with a therapy used at various times: a quart of cider a day, two spoonfuls of vinegar three times a day, a regimen of sea water, etc. One set of two was fed two oranges and a lemon every day. To quote: "The consequences were that the most sudden and visible good effects were perceived from the use of the oranges and lemons; one of those who had taken them, being at the end of six days fit for duty. The other was the best recovered of any in his condition; and being now deemed pretty well, was appointed nurse to the rest of the sick." It is of interest that it took the British Navy 40 years to supply oranges, lemon and lime juice (hence "limey") to its sailors at sea. That was how long it took for the Lords of the Admiralty to accept Lind's results. One can imagine the goings-on in the Advisory Councils to the Admiralty at the time; Lind was a misguided clinician, he had no statistical advice; how can you arrive at a definitive conclusion with such small numbers. The only thing missing is the recommendation that they needed more evidence – another trial with much larger numbers, which if performed would have eliminated scurvy in the British Navy some 39 years earlier.

Lind was primarily a clinician and there is no record of his having theorized on the method. However, Pierre Charles-Alexandre Louis (1789-1872), a physician, teacher, and chief of a hospital in Paris, not only conducted trials but also set down principles (Lilienfeld and Lilienfeld 1980). According to Lambert (1978), "In 1835 Louis tested the effects of bloodletting upon 78 cases of pneumonia, 33 cases of eryaipelas and 23 cases of inflammation of the throat. He found no appreciable differences in mortality or in duration of disease between patients bled or not bled, nor between those bled at different stages of the diseases. This result was so contrary to orthodox teaching of the time that it caused an uproar among his colleagues in Paris. Even Louis was reluctant to accept his own results and continued to treat certain diseases with bloodletting. Nevertheless, his observation led to a decreased use of this unhappy form of therapy." Louis later wrote on the "Numerical Method" in the assessment of treatments. Again I quote from his work (as given by Lambert): "As to different methods of treatment, if it is possible for us to assure ourselves of the superiority of one or other among them in any disease whatever, having regard to the different circumstances of age, sex, temperament, of strength and weakness, it is doubtless done by inquiring if under these circumstances a greater number of individuals have been cured by one means or another. Here again it is necessary to count. And it is, in great part at least, because hitherto this method has been not at all, or rarely employed, that the science of therapeutics is still so uncertain; that when the application of the means placed in our hands is useful we do not know the bounds of this utility. In order that the calculation may lead to useful or true results it is not sufficient to take account of the modifying powers of the individual; it is also necessary to know with precision at what period of the disease the treatment has been commenced; and especially we ought to know the natural progress of the disease, in all its degrees, when it is abandoned to itself." Skyrock (1947), an American medical historian, called this the introduction of mathematical procedures into clinical medicine.

In the l840's and 50's Elisha Bartlett, an American physician and a student of Louis, essentially formulated most of the basic tenets of the modern controlled trial, saying: "In therapeutic investigations cases which are to be compared must have equal disturbing factors of location, social class, and the like; they should be susceptible of a clear and positive diagnosis; there must be no selection of cases; and the method of treatment must be clearly defined. The certainty of results will be in proportion to the fixed and uniform character of the compared facts and to the greatness of their numbers –no acquaintance, however perfect, with laws of pathology and therapeutics, can ever remove, or in any degree diminish, the necessity of a thorough and discriminating study and knowledge of the single instances which unite to make up the materials of the law." One last reference to history. Lambert points out that a number of physicians and surgeons of the 19th century including Lister used Louis' method of comparing counting and meticulous observation. I quote "comparisons were nearly always made with a series of patients cared for prior to the new method of treatment. [What we now call historical controls.] This approach is effective only when the differences are great or when applied to highly fatal diseases such as meningitis or diabetes. No concurrent controls were required to prove the effectiveness of penicillin or insulin in these diseases."

The Problem of Selection Bias

Before proceeding to discuss the research I believe is needed in observational studies, let me give a few basic references. These include, a 1956 paper by H. Wold (1956) on "Causal Inference from Observational Data"; Cochran's paper (1965) on "The Planning of Observational Studies of Human Populations"; an excellent review by Sonia McKinlay (1975) on the "Design and Analysis of the Observational Study" with an excellent set of references; and the book, already referred to *Statistical Methods for Comparative Studies* (Anderson et al. 1980).

Now to get to the topic on which Dr. Colton wanted me to expand – the need for research on selection bias in observational studies. First let me clarify the topic itself. In my view, the research conducted by Cochran, Rubin and others on removal of bias refers to observed imbalances in baseline characteristics such as age, sex, disease severity if measured, and so on. Most of these studies concentrate on different methods for handling bias such as pairing, covariance, etc. In my view, the biases referred to are not one-to-one with selection bias. Recall I am only concerned with a comparative trial of therapies without randomization. How could this be planned prospectively? We choose a community and starting at a given date we collect information from all physicians using treatments A and/or B in their offices or clinics or hospitals. If this can be planned prospectively we may have a protocol including interview forms on admission and for all subsequent follow-up periods. If not planned prospectively then the study is limited to the data being collected. Again let me emphasize I am not advocating how observational studies should be done. I point to these approaches only to put some meaning to the concept of selection bias. In these circumstances, we all agree the serious barriers to valid conclusions is the patient self-selection bias and the physician selection bias. What do these mean? Has anyone ever seen any careful, precise, well spelled-out definitions of these terms? I suppose we can do that rather easily for the physician selection bias. He or she uses certain characteristics of the patient such as severity of disease, possibly age, duration of disease, condition of the patient and all other relevant factors in choosing a treatment for that patient. Or if the physician is only using one preferred treatment, he will use that information to select only those patients who he thinks have an increased probability of success on this preferred treatment.

As for the patient self-selection bias, I am not sure how this can be defined. If the patient goes to the clinician using treatment A because of physician traits – younger, older, good looking, etc. – how does this lead to bias? If the patient goes to that physician because he's using treatment A and treatment A is new this may constitute a selection bias. The patient may have already been on a number of past therapies without success. Does this increase or decrease the probability of success with treatment A? In the former we have a bias in favor of treatment A; the latter, a bias against treatment A.

Now no matter how these two concepts are defined, it is clear to me at least, that what is being stated is that each patient has his own probability parameter of success; that this probability differs from patient to patient and in fact there is a probability distribution of this parameter in the population of patients; that under the null hypothesis this distribution is the same for treatment A and B; and that a selection bias exists when different right tail areas of the distribution are represented by patients treated by A and patients treated by B: e.g., 60% of cases treated by A have success probabilities greater than 0.60. For example and very crudely, assume the distribution of the success probability among all patients is Beta with $\alpha = \beta = 2$. Then the distribution in the population is symmetric around a mean of $p = 0.5$ and 35% of patients have success probabilities greater than 0.60. Under random allocation, we would expect percentages of patients having a success probability greater than a given volume to be the same in the two groups. Under selection bias these percentages will differ.

This is all very nice and sounds good but what can be done with this formulation? There is a tremendous practical difference between these postulated success probabilities and baseline variables. The latter are measurable, the former are not. The interesting problem then is to see whether individual success probability parameters can be estimated or at the very least the average of the distribution of the success probabilities in each of the two treatment groups. Max Halperin (1985) suggests a procedure based solely on fitting the multiple logistic separately in each treatment group. In the treated group he would omit the treatment variable and let the treatment effect be absorbed in the intercept α. One would then compare the vector of coefficients of all the covariates – baseline, risk factors, etc. – between the two groups. If the vectors differ then there is some indication of the existence of some selection bias. The problem with this approach is the assumption that the selection can be specified in known covariates. Thus, it is possible for the two vectors to be the same yet a selection bias may still be operating.

If selection factors can be identified and measured then another possible research method to determine their effect and even adjust observed treatment differences could be as follows. Consider the successes as controls and the failures as cases and do a case control study in the usual way where the exposure variable is the suspect selection factor. Or one could compute a multiple logistic formation of several suspect selection factors, one for treatment A and another for treatment B. Someone has pointed out that Norman Breslow presented a procedure which may be very close to this if not the same at last year's annual meeting of the ASA in Detroit.

Lastly, research on selection is possible and perhaps may be more fruitful from another more empirical approach. There have been randomized clinical trials where the randomization procedure was not ideal and could have led to a selection bias. For example, a major randomized controlled clinical trial was initiated in 1948 to evaluate anticoagulant therapy in the prevention of coronary heart disease and myocardial infarction. There were 432 patients in the treated group and 368 in the control group. Groups were comparable with respect to age, sex, history of previous M.I.'s and severity of present attacks. Treatment, except for the anticoagulant therapy was similar in both groups. The results indicated the use of anticoagulants reduced the death rate, the number of new infarctions and the evidence of complications resulting from thrombi such as strokes and pulmonary emboli. The trouble with the study was in the randomization procedure and the absence of a double blind. Patients admitted on odd days were assigned to the treatment group and those admitted on even days to the control group. The referring and admitting physicians thus knew in advance which treatment their patients would receive and by manipulating admissions they could choose for their patients the form of therapy which they believed would be more beneficial. The potential for a selection bias would appear to be great. If one could have access to the original individual data, it would then be possible to try various schemes either to identify and estimate the selection effect or to compute adjusted treatment differences. For the former, one might analyze the treatment difference in mortality say for patients entered early and compare this with the treatment difference for patients entered later, the assumption being that some time would have elapsed for the physicians to start changing admission times. For the latter, one could match controls with cases in various ways to determine the effect on treatment differences. At this point, I suggest possible approaches to explore the effects of selection bias. It is not clear to me that any will turn out to be useful. But this is only a beginning. Such research could also be of help in resolving randomized clinical trials when the randomization scheme is suspect. It is obvious that if there are serious imbalances in observable baseline variables, it can only be because clinicians were manipulating patient assignment to a treatment. This by definition should give rise to a selection bias.

Let me conclude with the observation that statisticians must advise random allocation with concurrent controls when they are part of a group to plan the comparative evaluation of therapies or of the assessment of efficacy of one treatment. I think everyone agrees on this. But we cannot shrug our shoulders and advise not to do any study when conditions are such that random allocation cannot be done or when concurrent controls are clearly not necessary. Finally, we must remember that although randomized control trials make it more likely for our inferences to be valid and non-randomized studies are less likely for inferences to be valid, there are *no* theorems which state that observational studies cannot yield valid inferences.

References

Anderson S, Auquier A, Hauck WW, Oakes D, Vandaele W, Weisberg HI (1980). *Statistical Methods for Comparative Studies: Techniques for Bias Reduction*. New York: John Wiley & Sons, Inc.

Cochran WG (1963). *Sampling Techniques, 2nd ed.*. New York: Wiley & Sons, Inc.

Cochran WG (1965). The planning of observational studies of human populations. *J Royal Statistical Society, Series A*, 128:234-265.

Cochran WG and Rubin DB (1973). Controlling bias in observational studies: A Review. *Sankhya, Series A*, 35:417-446.

Cornfield J (1966a). Sequential trials, sequential analysis, and the likelihood principle. *American Statistician*, 20:18-22.

Cornfield J (1966b). A Bayesian test of some classical hypotheses with applications to sequential clinical trials. *Journal American Statistical Association*, 61: 577-594.

Cornfield J (1969). The Bayesian outlook and its applications. *Biometrics*, 25:617-657.

Cutler, SJ, Greenhouse SW, Cornfield J, and Schneiderman MA (1966). The role of hypothesis testing in clinical trials. *Journal of Chronic Disease*, 19:857-882.

Greenhouse SW (1980). Some epidemiologic issues for the 1980's. *American Journal of Epidemiology*, 2:269-273

Halperin M, Abbott RD. Blackwelder WC, Jacobowitz R, Lan G, Verter J, and Wedel, H (1985). On the use of the logistic model in prospective studies. *Statistics in Medicine*, 4:227-235.

Hill, A.B. (1949). *Principles of Medical Statistics*. London: The Lancet Limited.

Hinkley, David V. (1983). Can frequentist inferences be very wrong? A conditional "yes". In *Scientific inference, data analysis, and robustness*, p. 85-104

Lambert EC (1978). *Modern Medical Mistakes.* Bloomington, IN: Indiana Univ Press.

Lilienfeld DE and Lilienfeld AM (1980). The French influence on the development of epidemiology. *Bull. History of Medicine*.

Lind, J. (1753). *A Treatise of the Scurvy.* Edinburgh: Sands, Murray, and Cochran.

McKinlay, S M (1975). The design and analysis of the observational study – A review. *J American Statistical Association*, 70: 503-520

Shyrock RH (1947). *The Development of Modern Medicine*. New York: Knopf.

Wold H (1956). Causal inference from obervational data (with discussion). *J. Royal Statistical Society, Series A*, 119:28-60.