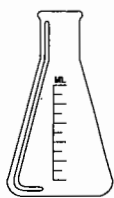


10

## Biopharmaceutical Section



American Statistical Association

# Biopharmaceutical Report

Volume 1, No. 3

Fall 1992

Chair: *Camilla Brooks, Ph.D.*

Editor: *Avital Cnaan, Ph.D.*

## Safety Analysis: Too Much? Not Enough? and How?

**Christy Chuang-Stein**

*The Upjohn Company*

### Introduction

No one can deny the importance of collecting safety data in clinical trials, especially trials involving an investigational drug or a novel treatment. The overwhelming concern for a patient's safety in a trial creates an industry that can spend as much as 70% to 80% of its time and effort to collect data entirely from a safety perspective. Safety data ranging from clinical signs/symptoms, safety laboratory assays to physiological tests such as ECG are routinely recorded on patients' case report forms. A pharmaceutical sponsor typically has well-established in-house standard operating procedures to report adverse experiences encountered in a trial. Because of the fear of being amiss at important adverse experiences, investigators are generally encouraged to over-report medical events in a trial rather than under-report them, contributing to the amorphous and diverse nature of the safety data collection.

In addition to assuring a patient's safety in a trial, safety data allows one to study the safety profile of a drug or treatment. For this purpose, safety experience is compiled from each individual in a trial and is frequently aggregated across studies. Pooling the safety experience from different studies is especially important because most clinical trials are designed to achieve objectives in efficacy and very few of them individually have adequate power to evaluate safety.

The need to summarize and analyze safety data has undoubtedly spun a flurry of research, or at least writing, on this subject. A detailed discussion on methods available to study different safety endpoints for various types of trials was given by O'Neill (1988). Among statisticians, opinion regarding how safety data should be analyzed varies greatly. While some support the use of inferential statistical methods (Enas, 1991), others stay away from them (Huster, 1991). Many statisticians take the middle ground and use the p-values obtained from inferential statistical methods for exploratory purposes (Abt, 1987, 1990). In terms of analyzing lab data, Sogliero-Gilbert, Mosher and Subkoff (1986) proposed a method to combine related lab results to study the extent of functional impairment reflected by the lab results. Similar to Sogliero-Gilbert, Mosher and Subkoff's idea of examining several parameters at a time, a vector-based re-sampling method was employed by Rampey and Enas (1991) to study a cluster of adverse events within a body system. Realizing that all of the above efforts concentrate on one-kind-at-a-time analyses and safety data are really multi-faceted, Chuang-Stein, Mohberg and Musselman (1992) propose to combine all relevant safety data, organize them by body systems and analyze the consolidated information using a multivariate approach.

What constitutes an appropriate safety analysis for a given trial with its unique objectives? Are we conducting too many comparisons, or is what we are doing (if we

## Contents

### FEATURED ARTICLE

#### Safety Analysis: Too Much? Not Enough? and How?

..... CHUANG-STEIN 1

Discussion ..... ENAS 4

Discussion ..... ROBERTS 6

Discussion ..... SALSBURG 9

Discussion ..... SRINIVASAN 9

Discussion ..... WITTES 10

### BIOPHARMACEUTICAL SECTION NEWS

Letter from the Editor ..... 11

Journal Response: Letter to *Clinical Pharmacology and Therapeutics* ..... 12

Book Review: *Design and Analysis of Bioavailability and Bioequivalence Studies* ..... 13

Software Review: S-Plus ..... 15

1993 Biopharmaceutical Executive Committee ..... 18

Section-Sponsored Sessions at 1993 ENAR Meetings ..... 20

### CONFERENCE REPORTS

Section-Sponsored Sessions at 1992 Joint Statistical Meetings ..... 13

know what we are doing) enough? Are we doing the *right* thing? Have we summarized results from clinical trials in the best possible way to help a practicing physician understand the safety outlook of a new treatment and facilitate the treatment selection when safety is the primary consideration? I have often asked myself these questions. The answers, it appears, should be sought with a mind broader than what is currently driving the safety analysis.

To help the discussion in this paper, I will focus on several issues associated with the safety analysis. I will start by examining the different needs of safety analyses at different stages of drug development or treatment evaluation, and point out some deficiencies of the current way of conducting safety analyses. Some suggestions and recommendations will, hopefully, stir further discussion on this topic. I hope this article provokes more thoughts and leads to a more in-depth examination of what we are currently doing and how we can possibly do a better job in summarizing and analyzing the safety data. Because of space limitations, an extensive literature review is omitted from this paper.

### Objectives of Safety Analyses

For convenience, I will use the term "safety analyses" to mean the summarization and compilation of the safety experience of individuals participating in a trial or trials involving an investigational drug or a novel treatment. Even though each individual's experience in a trial is important by its own account, it is the aggregate of such experience that helps to identify the potential safety concerns of a new treatment modality when it is given to a target patient population.

#### 1. Phase I/II

The primary objective of a safety analysis in phase I/II clinical trials is to identify the most frequent side effects of a new drug/treatment and to study its overall safety profile. Because of the limited study size, we are likely to observe only those side effects that have a relatively high incidence rate. The typical frequency listing of adverse events with their intensities plus a separate analysis of the lab parameters generally suffice for the identification of the most frequent side effects. Nevertheless, these separate summarizations do not help to bring together findings from separate analyses to produce an overall safety profile. To produce the overall safety profile, one needs to resort to all safety information and combine such information from all sources. This need exists for all pre-marketing trials, and especially so for phase I/II clinical trials when the experience with the drug is limited and one is not sure what to look for.

Since safety information comes from multiple sources that can be equally important, the simultaneous analysis of such information resembles that of the efficacy data when there exist multiple efficacy endpoints. While multiple efficacy endpoints have received a sizeable attention (Follmann, Wittes, and Cutler, 1992; O'Brien, 1984; Pocock, Geller, and Tsiatis, 1987; Gelber, Gelman, and Goldhirsch, 1989, etc.), multiple safety endpoints have not. Chuang-Stein, Mohberg and Musselman (1992) proposed to structure the massive safety data into a more manageable framework by consolidating them into a number of classes characterized by body systems and determined in conjunction with the underlying disease as well as the treatment(s) involved. Within each class, they propose to assign to each patient an overall intensity grade based on all relevant information. The analysis of such organized data concentrates on a simultaneous comparison of the mean intensity grades for different treatments within each class with the use of a multivariate statistic and scores that reflect the acceptability of the various intensity levels. One advantage of their approach is its ability to utilize all pertinent safety information to come up with a vector of intensity grades that reflect an individual's overall experience within the various designated classes. The trade-off is the decision rule that one needs

to determine beforehand to consolidate the information. On the other hand, the decision rule forces one to think hard in advance about what constitutes a more serious safety concern for a target population. Such an exercise also mimics the practice that a physician needs to go through when presented with disjoint safety summaries.

#### 2. Phase III

Because of their size and randomized nature, phase III trials provide the best pre-marketing safety data. Typical safety analyses of phase III trials consist of comparing the distributions of the medical events (with and without the associated intensities) between the two treatment groups. At times, analysis by hierarchy of events is conducted. Such analysis examines the occurrences of the following in sequence: (i) death due to medical events; (ii) death and hospitalization due to medical events; (iii) death, hospitalization and dropout due to medical events; (iv) death, hospitalization, dropout and dose modification due to medical events; (v) individuals experiencing any medical event. As for laboratory parameters, analysis of variance (or covariance) is frequently applied to the change in these parameters. Other event occurrences, such as abnormality observed on ECG or chest X-ray, are typically compared using binary outcome techniques (with or without adjustment for covariates).

All the above analyses are informative in their respective roles. There are various approaches to handle the multitude of inference (in particular, the p-values) associated with the analyses. Some people do not conduct any formal statistical comparisons on the majority of the safety data while others declare a significant difference between the treatment groups only if the number of p-values less than 5% is less than the number of total comparisons times 5%. Still, some people choose to interpret only those p-values that are extremely small. There does not seem to exist a universal rule on how the rich field of safety data should be plowed. Worse yet, many the comparisons appear to be data-driven. The latter partially results from FDA's (1988) requirement that rigorous statistical methods be applied only to events with substantial differences that are potentially useful to prescribing physicians.

In addition to not fully utilizing safety data as discussed above in the context of phase I/II trials, safety analyses for phase III trials carry a different mission. Since phase III trials are the formal arena where treatments are compared directly, they provide the essential information for treatment selection based on both efficacy and safety consideration. As a result, safety analyses for such trials go beyond the identification or a mere summarization of adverse experiences.

Interventions are given with the hope to cure, or to control, existing disease or to palliate discomfort and pain. Along with the benefit, there is also the danger that an intervention can cause harm or injury. Depending on the disease or symptom that an intervention is developed to treat, the associated harm might not be acceptable. This is particularly so with diseases which, although may be a cause of considerable discomfort, are not life-threatening. In these cases, tolerance for treatment-induced harm are low. On the other hand tolerance for treatment-induced harm is high for life-threatening diseases such as cancer and AIDS. This intertwined relationship between benefit and risk has led to the general belief that risks should be evaluated with respect to the achievable benefit and that benefit-risk assessment should be made with respect to the underlying disease or symptom.

Despite the recognition of the inseparable roles of benefit and risk in the evaluation of treatments, the analysis of clinical trials still lacks a general formulation of a benefit/risk assessment with risk measured by the safety data. The pharmaceutical industry routinely prepares separate efficacy and safety summaries in a new drug application, and it is uncommon that a serious effort is launched in the same application to include the two endpoints in a joint analysis. As a result, if one treatment provides more efficacy

with a price tag of more side effects, the decision on the treatment selection becomes a difficult one. Some formal statistical tools are definitely needed to address the simultaneous benefit/risk comparisons between treatments. Chuang-Stein, Mohberg and Sinkhole (1991) made an attempt in this direction using ordinal response data. Statisticians need to give more thought to the practice of merging efficacy and safety in one analysis with input from the medical personnel.

### 3. Post-Marketing

Safety analysis for post-marketing trials is an entirely different issue from that for pre-marketing trials. Some complicating factors were discussed in detail by O'Neill (1988). Because some side effects do not surface until much later, post-marketing surveillance can pick up side effects that were not observed during the pre-marketing testing. An example is the incidence of somatic and genetic deaths following mass chest X-rays exams to pick up tuberculosis, lung cancer and leukemia as discussed in Paine and Liken (1975). In addition, rare side effects start to show up during the post-marketing epidemiologic follow-up studies. I highly recommend conducting additional benefit/risk assessments in the presence of newly detected side effects after a new drug/treatment has entered the market.

### Integrated Safety Summary

Because of FDA's mandate on an integrated safety summary for a new drug application, the sponsor of a new drug or a novel treatment religiously pools data from different studies to create the required safety summary. A typical approach at the moment is to collapse data from all studies and summarize them as if they came from a single study. But, is this approach appropriate? Differences in treatment plans and target populations can introduce extra variation to the parameters of interest, but the current way to construct the integrated safety summary does not include a provision for such a variance component. Even in the absence of the FDA's mandate on the integrated safety summary, safety data should be pooled from different studies to increase our experience with the safety outlook of the new drug or the treatment. The question is: How should the pooling be conducted?

Combining results from different studies has received a fair amount of attention since the 1980's. The methodology generally comes under the title of meta-analysis. Meta-analysis pools results from different studies while recognizing the differences among the studies. Even though meta-analysis has been employed mostly for efficacy evaluation, this type of analysis has a role in pooling safety data from different studies as well. In particular, the two-stage sampling approach assuming a prior distribution for the parameters of interest is flexible enough to enjoy a wide range of applications. This approach incorporates a variance component that addresses the fact that data may exhibit a between-study variability. The extent of the variability will be estimated from the data. Therefore, if the variability among the observed summary statistics from different studies is low, the distribution assumed for the parameters will be estimated to be nearly degenerate.

When pooling data from different studies, one needs to take into account the objective of pooling. If one is interested in the safety profile of an older patient population with impaired liver function, one should pool data from only those studies with such a target patient population. Other considerations include the dose received and the exposure duration.

### Use of Reference Ranges

Without dispute, laboratory results are the most reliable indicators of systemic toxicities and provide vital information regarding a patient's safety in a trial. The interpretation of laboratory results is commonly done by utilization of a reference range. This practice arose from the need to identify *diagnostically*

useful deviations in laboratory measurements. The use of lab-specific reference ranges is especially common in multi-center trials where patient's specimens are being processed at the respective study sites for clinical interpretation. The use of lab-specific reference ranges creates a belief that as long as the clinical interpretation of lab results is done using lab-specific ranges, the interpretation is sound. Nevertheless, the use of lab-specific reference ranges which are based on *presumed* normal population misses one fundamental issue, i.e., whether the new drug/treatment has adversely altered a patient's pre-existing biochemistry in the target patient population. This becomes a larger problem for trials involving advanced cancer or AIDS patients who frequently present with multiple lab abnormalities prior to receiving any of the treatments studied in a trial.

There are at least two basic problems with the use of reference ranges to interpret lab results in the context of monitoring a patient's safety in a clinical trial. First, most of the laboratories participating in a clinical trial use assays intended for diagnostic evaluations of a disease state, and not for assuring the continuation of a pre-existing status of a patient's biochemistry. Unfortunately, the latter is what safety monitoring in a clinical trial should focus on, except when the trial enrolls only normal volunteers or patients who are basically healthy. A more serious problem, which we as statisticians are generally unaware of, is that there is no universal definition for the reference population and there is no consensus regarding the method to determine the reference ranges. The heterogeneity in the selection of the reference population and the method to construct a reference range casts some doubt on the usefulness of the reference ranges, especially in a trial involving multiple laboratories.

Other sources that contribute to the variability in reference ranges include differences in the assay procedures and the accuracy in diagnosis. First, even though analytical accuracy is a desirable feature of an assay, it is an ill-defined target. When setting up an assay, a laboratory must take into account the clinical intentions of its users and select the operating characteristics of an assay accordingly. Because lab assays are traditionally used to diagnose a disease state, a laboratory tends to balance the reagents in such a way that the resulting biochemistry is the most precise at or near the point where clinical decisions are to be made. These usually are at the upper and lower ends of the reference population ranges. On the other hand, clinical trial laboratories must optimize precision so that the precision of the results outside the reference range are equal, or even better than the precision at the upper and lower ends of the range. An example is the insulin and C-peptide levels that are used for assessing the performance of anti-diabetic medications.

Reference range was once called *normal* range. The implication is that a reference range defines a region of lab results within which the likelihood of there being a biochemical abnormality in a patient is relatively small. Unfortunately, the analysis level for many disease entities is very close to, or even overlaps, the range of values observed in a normal healthy population. As a result, the determination of the range is also influenced by two questions: "What degree of confidence does one want that a result outside the reference range is indeed abnormal (sensitivity)", and "What degree of confidence does one want that a result inside the reference range is indeed normal (specificity)"? Which question carries more weight depends, again, on a laboratory's clientele. Thus, even if the analytical procedures are identical, the concerns about the rates of false positive and false negative can lead to different ranges. Instead of using reference ranges obtained from individuals who are presumed to be normal, Oliver and Chuang-Stein (1993) proposed to use a study's inclusion/exclusion criteria to define a reference population. In addition, they recommended using patient's pre-treatment lab results adjusted by a laboratory's performance in a proficiency survey to construct a set of study-

specific reference ranges, instead of using lab-specific reference ranges to determine lab abnormality.

A proficiency survey is a quality-control program which is intended to serve as a means to identify aberrant lab performance by having all enrolled laboratories analyze a common set of specimens. After analyzing the specimens, laboratories send their results to one coordinating center which collates and groups the results by assay methodology for an "apples-to-apples" comparison. The laboratories are required to run the test samples in the same manner as they analyze patient's specimens to ensure that the proficiency testing gives a true picture of the performance expected of the laboratories when they run samples for clinical interpretation. Oliver and Chuang-Stein (1993) recommended a two-stage adjustment procedure that adjusts lab values based on a laboratory's performance observed in the proficiency survey. The advantage of the suggested adjustment comes from the fact that reference ranges that are more relevant to the study patient population are being used to monitor patient's biochemical status after the onset of the treatment. Also, lab values are adjusted based on their performance on a common set of samples. Such an adjustment facilitates the summarization of data from different laboratories within one study where differences occur as a result of laboratories using different procedures or equipments.

When analyzing lab data, it is a common practice to generate a table that flags lab values (at all evaluation points) that are outside the reference ranges with the ranges being those supplied by individual participating laboratories. Since the current reference ranges are constructed based on data from a *presumed* normal population, this practice does not make much sense for situations where patients enter a trial with severe functional impairment. In addition, this practice infers nothing on how a patient's biochemistry was affected by the treatment. But then, why are we still doing it and why is it still being routinely required by the regulatory agency?

### Other Considerations

Because of the diverse nature of safety data that are routinely collected in a trial, the techniques to summarize and analyze them are less obvious and straightforward compared to those for the efficacy data with one exception. The need to conduct a benefit/risk assessment is not unique to the safety analysis; it is applicable to the efficacy analysis as well. A safety analysis should be such that an overall safety profile can be drawn based on data from all sources. This requires an effort to intelligently combine all safety information to facilitate the overall summarization. The effort takes work, but the work enables one to put the puzzle pieces together to create a more succinct picture of the safety outlook of a new treatment.

One thing that I have not touched upon but nevertheless requires an equal amount of thought, is the collection of the safety data itself. Because there usually do not exist any definite rules governing the collection of medical events, their collection is both amorphous and irregular. Before starting a trial, one should think hard how the medical events should be collected. Should medical events be solicited open-endedly or using a checklist? Should they be volunteered? How should multiple episodes of the same medical events be handled? How should a change in the intensity of a clinical event be recorded? Fairly little attention has been given over time to some of these issues. If we are going to spend so much effort and time in a trial to collect safety data, shouldn't we do it as consistently as we possibly can and make the best use the data allows us?

### References

Abt K (1987). Descriptive data analysis: a concept between confirmatory and exploratory data analysis. *Methods of Information in Medicine*, 26, 77-88.

- Abt K (1990). Statistical aspects of neurophysiologic topography. *Journal of Clinical Neurophysiology*, 7, 519-534.
- Chuang-Stein C, Mohberg NR, and Sinkhole MS (1991). Three measures for simultaneously evaluating benefits and risks using categorical data from clinical trials. *Statistics in Medicine*, 10, 1349-1359.
- Chuang-Stein C, Mohberg NR, and Musselman, DM (1992). Organization and analysis of safety data using a multivariate approach. *Statistics in Medicine*, 11, 1075-1089.
- Enas GG (1991). Making decisions about safety in clinical trials - the case for inferential statistics. *Drug Information Journal*, 25, 439-446.
- Follmann D, Wittes J, and Cutler JA (1992). The use of subjective rankings in clinical trials with an application to cardiovascular disease. *Statistics in Medicine*, 11, 427-437.
- Gelber RD, Gelma RS, and Goldhirsch A (1989). A quality-of-life oriented endpoint for comparing treatments. *Biometrics*, 45, 781-795.
- Guideline for the Format and Content of the Clinical and Statistical Sections of an Application*, Center for Drug Evaluation and Research, Food and Drug Administration, Department of Health and Human Services.
- Huster WJ (1991). Clinical trial adverse events: the case for descriptive techniques. *Drug Information Journal*, 25, 447-456.
- O'Brien PC (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, 40, 1079-1087.
- Oliver LK, and Chuang-Stein C (1993). Laboratory data in multicenter trials: monitoring, adjustment and summarization. A chapter to be included in *Drug Safety Assessment in Clinical Trials* (ed. Gene Gilbert). New York: Marcel Dekker, Inc.
- O'Neill RT (1988). Assessment of safety. Chapter 13 in *Biopharmaceutical Statistics for Drug Development* (ed. Karl E. Peace). New York: Marcel Dekker, Inc.
- Paine JT, and Likens MK (1975). A survey of the benefits and risks in the practice of radiology. *CRC Critical Reviews in Clinical Radiology and Nuclear Medicine*, 6, 425-439.
- Pocock SJ, Geller NL, and Tsatis AA (1987). The analysis of multiple endpoints in clinical trials. *Biometrics*, 43, 487-498.
- Rampe AH, and Enas GG (1991). Current issues and innovative approaches in adverse event data analysis. Presented at the 1991 Drug Information Association Meeting, Washington DC.
- Sogliero-Gilbert G, Mosher K, and Zubkoff L (1986). A procedure for the simplification and assessment of lab parameters in clinical trials. *Drug Information Journal*, 20, 279-296.

## Discussion

### Quality Safety Information that Meets Customers Needs

Gregory G. Enas

Lilly Research Laboratories, Eli Lilly and Company

Dr. Christy Chuang-Stein is to be applauded for all of her recent work in making us think harder about analyzing safety data in clinical trials. Her present work is a nice synopsis of many of the current issues and has helped to consolidate and solidify my own thinking about these dilemmas. In a nutshell, the questions of Too much?, Not Enough? and How? might be better addressed if we first seek answers to **Who is the Customer?** and **What Design?** I know that statisticians have not gotten involved much in the analysis of safety data until recently. Since the collection of safety data has often been accomplished with little design in mind, statistical thinking may have been an afterthought on the analysis side.

As Christy has ably pointed out, safety is truly a multivariate characterization which often is the most ill-defined quantity in the benefit/risk ratio. Her previously noted work on consolidating safety data into classes is a practical attempt to treat safety as a multivariate problem. Christy's other work on realistic laboratory reference ranges and focus on changes with respect to the reference range is also an example of how to better define a compound's safety profile. The hierarchical analysis of events she describes is intriguing and appears to be adaptable to important events other than those specifically mentioned. For example, administration of concomitant medication to ameliorate drug toxicity could also be factored in the cascade of events.

### **Who is the Customer?**

To really get a handle on analysis, however, we have to know who our customers are and what are their needs. Christy mentions the FDA and regulatory agencies in general as one basic customer. She takes note of many of the initiatives that have been undertaken to address the needs of regulators. The integrated safety summary is certainly a very important information piece around which much activity has been generated. For example, standardized output tables and analyses are being developed so that review and comprehension of a safety data package is maximized. Not only are output tables and analyses being standardized but the data generation process is being standardized. Central laboratories are becoming more prominent because of the need to pool laboratory data from many different investigative sites. Use of central laboratories to handle all types of patients with many kinds of medical conditions facilitates creation of reference ranges that are meaningful instead of reliance on "normal" ranges. Common adverse event dictionaries and standard event outcome measures, such as Treatment Emergent Signs and Symptoms (TESS), also help expedite review of the safety profile of a new treatment (Offen, 1988). Many of these innovations we have seen come to pass for premarketing safety analyses may play an important role in the expansion of post-marketing surveillance studies and prospective controlled studies.

Another key customer are the investigators who conduct the studies. In particular, much of the premarketing clinical safety experience is derived from studies primarily designed to demonstrate efficacy. As Christy points out, many of the early clinical studies are only able to identify the most frequent adverse events possibly associated with a new treatment. We have found that investigators find safety data monitoring rules very useful tools in these early studies where one does not know what to expect or only has a limited hunch based on animal toxicology. These rules take the investigators' safety concerns and uncertainties and translate them into counting processes which reflect the degree to which the investigator is willing to continue to treat patients in a study given the cumulative evidence for treatment-related toxicity (Wiltse, Eras, et al, 1993). Once events have been identified that need very careful monitoring, these rules and other Bayesian or classical sequential monitoring schemes help give investigators objective assurance that safety issues in the conduct of a study are being addressed in a timely manner.

Yet another primary customer are the health care providers, primarily physicians, and the patients themselves. As O'Neill (1988) has pointed out, drug labeling is a primary mechanism for informing the physician and patient about a product's benefits and risks. The package insert can contain so much information that one may find it very difficult to feel comfortable about a product without much experience with it. Some of this information overload may also be due to another important consumer of safety information, namely those trained to deal with the legal aspects of drug development and marketing. Adverse events reported in <1% of treated patients, regardless of the relationship to treatment, are listed and may number in the hundreds. Laboratory changes

occurring in <0.1% of patients may also be listed, regardless of treatment relationship. At issue here is the fact that all safety information must be summarized as succinctly as possible yet without obfuscating any individual adverse events, drug interactions, or patient prognostic characteristics (e.g., elderly versus non-elderly patients, photosensitivity). Graham (1991) and Litka (1991) address various labeling issues.

### **What Design?**

The customers dictate what information is useful to them. Statisticians can help meet the varied needs of these customers by addressing issues of design. The above example of safety monitoring rules shows how a statistician can help a physician monitor a study by collecting and analyzing safety information on an ongoing basis throughout a study. In a similar manner, statisticians can continue to improve their working relationships with study physicians and monitors, FDA scientists, and other regulatory personnel in order to standardize collection, analysis, and reporting of safety data. In many therapeutic areas for example, a much larger safety database is necessary for approval even though efficacy might be demonstrated in a smaller number of patients. Why not design these "safety" trials to actively solicit information on certain adverse clinical, laboratory, and other events that may have been spontaneously generated in early efficacy studies? Proper control is necessary to assess treatment relationships. Even confirmatory trials can take advantage of early safety information by designing them so that information on potentially primary safety issues is solicited.

I agree with Christy that the collection of safety data requires an equal amount of thought. In fact, it may require even more thought, because the design and collection of data dictates what the analysis will be and the strength of the inference. One should think hard about how they want the label to read. What adverse events, for example, should be listed in the package insert which allow one to distinguish a certain product or be primary in defining the benefit to risk ratio? These events can be solicited, especially in later Phase III type studies, with the degree of detail that is warranted. Intensity, duration, concomitant treatment, multiple occurrences, and other aspects of these particular events could be reliably quantified and give practical information to the prescribing physician if thought out a priori. The design of these studies could then take these events into account when establishing sample sizes and the standardization of safety data collection.

Christy mentions the integrated safety summary as an exercise where the pooling of data and subsequent analysis is not trivial. To take this one step further for the practicing physician, I have not seen evidence to date of package inserts being crafted from carefully conducted meta-analyses. For fun I reviewed all of the drug advertisements and associated package inserts in the October 15, 1992 issue of *The New England Journal of Medicine*. There were a few instances where some very large, well done safety studies were acknowledged in great detail, including risk ratios, p-values, and confidence intervals. For the most part however, especially in the Adverse Events section, the data was usually all pooled together without any comment on any differences in patient types studied, investigator differences, treatment by sub-type interactions, and the like. This is not to say that this pooling is inappropriate. It just doesn't give any evidence that we have thought long and hard about what and what not to pool and why.

Maybe if we begin to think hard about this when conducting an integrated safety summary, this will indeed translate, when necessary, into more informative package inserts. This will demand much more thinking about pooling all kinds of studies, controlled, uncontrolled, different doses, etc. in a rational manner. Maybe some more attention like that shown in Begg and Pilote (1992) would be helpful here. For events with relatively high frequency, meta-analytic thinking can help sort out different

patient subtypes and differences amongst therapies. For rarer events, meta-analysis helps sharpen the signal/noise ratio.

Recent work on Bayes Causality Assessment may prove useful in reducing the noise present in a safety database by collecting information that has a certain probability of treatment relatedness. Trial designers can plan to collect such information and base treatment comparisons on events that meet the causality criteria. If solicitation of important events becomes more prevalent, then one needs to determine whether to pool and how to pool unsolicited events, solicited events, TESS, treatment related events such as those determined using Bayes rules, and other definitions together. In this regard we should remember that patient therapy, and trial characteristics are often the most important issues to consider when assessing whether to pool or not. However, the need for some commonality of definition is still very strong.

Christy also mentions the importance of post-marketing surveillance studies. Some interesting ideas in this regard are given by Tsong (1992), Norwood and Sampson (1988), Rawson et al. (1990), and Barbujani and Calzolari (1984). Note the emphasis on using innovative quality-control measures as a means to monitor post-marketing safety data.

### Summary

Dr. Chuang-Stein's paper about the analysis of safety data leads us to also consider the important questions of who needs this information and what can be done proactively to get the right information to them? This means that statisticians must play stronger roles in the design of studies from a safety data perspective. This involves study designs which include ways to monitor safety data throughout the course of pre- and post-marketing studies. Statisticians should also be thinking about solicitation of key safety information rather than unsolicited collection in (Phase III) studies, use of realistic reference ranges for laboratory data, pooling different studies into integrated safety summary reports, and meta-analytic techniques.

Hopefully these and other measures will contribute to improved patient care through better informed medical practitioners, regulatory officials, and biopharmaceutical researchers.

### References

- Barbujani G and Calzolari E (1984). Comparison of two statistical techniques for the surveillance of birth defects through a Monte Carlo simulation, *Statistics in Medicine*, 3, 239-247.
- Begg C and Pilote L (1991). A Model for Incorporating Historical Controls into a Meta-Analysis, *Biometrics*, 47, 899-906.
- Graham G (1991). Labeling: What should it say, and how should it say it? *Drug Information Journal*, 25, 211-216.
- Litka P (1991). Labeling Development: Incorporation of safety information, *Drug Information Journal*, 25, 205-210.
- Norwood P and Sampson A (1988). A statistical methodology for postmarketing surveillance of adverse drug reaction reports, *Statistics in Medicine*, 7, 1023-1030.
- Offen W (1988). Statistical evaluation of adverse events, *ASA Proceedings of the Biopharmaceutical Section*, 1-8.
- O'Neill RT (1988). Assessment of safety, Chapter 13 in *Biopharmaceutical Statistics for Drug Development* (ed. Karl E. Peace). New York: Marcel Dekker, Inc.
- Rawson N, Pearce G, and Inman W (1990). Prescription event monitoring - methodology and recent progress, *Journal of Clinical Epidemiology*, 43, 509-522.
- Tsong Y (1992). False alarm rates of statistical methods used in determining increased frequency of reports on adverse drug reaction, *Journal of Biopharmaceutical Statistics*, 2, 9-30.
- Wiltse C, Enas G, Johns D, Brunelle R (In preparation). Defining early stopping rules for a clinical trial based on observed numbers of adverse events.

## Discussion

### Changing the Reporting Process of AEs

Robin S. Roberts

Department of Clinical Epidemiology and Biostatistics  
McMaster University

### Introduction

It is by no means an exaggeration to state, as does Dr. Christy Chuang-Stein, that 70-80% of the effort expended in data collection associated with clinical trials of new drugs concerns safety. Although fairly innocuous in appearance, the open-ended text-oriented case report form (CRF) sections dealing with adverse experiences (AEs) and concomitant medications are disliked by participating investigators, dreaded by company monitors, frustrate data managers, and present an analysis challenge to biostatisticians. While none of the trial participants questions the need to address safety, there is a barely disguised ground swell of opinion that much of this time consuming effort is uninformative. The relatively few important adverse reactions or drug interactions are at risk of being submerged in a sea of data collection noise, thereby causing us to "lose the baby with the bath water" at the time of analysis. Laboratory data, being inherently quantitative, creates fewer problems for the trial management but nonetheless can cause a severe case of biostatistician's heartburn (a recognized side-effect in most trials) when analysis time comes around.

The heart of the problem, I believe, is the inherent conservatism of many pharmaceutical companies. While one can understand their motivation, I sometimes wonder if they appreciate the potential negative consequences of their conservativeness. In protecting the public's well-being, the FDA essentially defines what it means by an AE and in addition what are serious and life threatening AEs. Companies are required to report promptly to the FDA all serious AEs which are also "unexpected and associated with the study drug." The terms "unexpected" and "associated" are further defined in the FDA guidelines, the latter being that there is "a reasonable possibility that the experience may have been caused by the drug". In practice, many companies require investigators to report all serious AEs quickly to a company clinical monitor irrespective of, for example, how likely they are to have been drug related. The clinical monitor then decides which to relay on to the FDA. In many studies which deal with chronically sick patients, only a small proportion (possibly only 1%) of serious AEs reported to the company may be passed on to the FDA.

The same conservatism spills over into the design of CRFs in that typically investigators are required to provide extensive documentation on everything bad that has happened to a patient. The CRF does not usually distinguish (in terms of information required) between events which have a "reasonable possibility" of being drug related (i.e. adverse reactions or side effects) from the plethora of other bad things that can happen to the sick patients recruited into trials. The net result is that we are asked to report (in what some might view as excruciating detail) many AEs which are almost certainly manifestations of the underlying disease process which qualified the patient for the study or are other unrelated comorbidities. This procedure must have potential costs in terms of the overall data quality and completeness of reporting as well as creating the "baby with the bath water" situation.

In my view, the companies' conservatism stems from a reluctance to believe that clinicians can reliably distinguish AEs which have a reasonable possibility of being caused by study medication from those due to an unrelated comorbidity. I am not suggesting that events deemed to be unrelated need not be reported, but that the reporting requirements be much reduced in

**Table 1: All Adverse Experiences - CATS**

Treatment Group	AEs	Relationship to Study Medication		
		Probably Not Related	Unknown	Probably Related
<b>PLACEBO</b>				
	Number Reports	127	156	93
	Rate/100 Assess.	4.1	5.0	3.0
<b>TICLOPIDINE</b>				
	Number Reports	176	334	275
	Rate /100 Assess.	6.4	12.1	9.9
Rate T/Rate P		1.6	2.4	3.3

terms of the amount of data requested. In this way the participating clinicians can concentrate their finite store of documentational energy into the adverse reaction corn rather than the unrelated comorbidity chaff.

In theory, I think it is relatively straightforward to determine whether clinicians are up to the task of recognizing adverse reactions from within the more generally defined set of adverse experiences. From our experience, at least, the answer is quite clearly yes. To support this position, I offer data from two recently completed studies coordinated by the group in which I work. The first, the Canadian American Ticlopidine Study (CATS) (Gent, 1989) was a placebo-controlled trial to assess the efficacy of the anti-platelet drug Ticlopidine in reducing the subsequent risk of vascular ischemic events in patients who had a stroke. A total of 1072 patients were recruited from 25 centers in North America and the study was conducted under US IND regulations. Patients were followed at essentially 4 month intervals for an average of 24 months. The AE section of the CRF was fairly typical and asked clinicians to judge the relationship to study drug as "probably related", "probably not related", or "unknown". Ticlopidine proved to be efficacious, but like many active drugs had "active" side effects. In Table 1, I have summarized the rate of all AE reports separately for placebo and Ticlopidine, subdivided by the likelihood of being drug related. The ratio of the frequency of reports was over 3 to 1 for the subgroup judged as "probably related", and only 1.6 to 1 for those judged "probably unrelated", with the "unknown" group falling in the middle with a ration of 2.4 to 1. Assuming the blinding was effective, this supports the

contention that clinicians can sort out real adverse reactions although the ratio of 1.6 for the unrelated group indicates that some are overlooked. Table 2 shows the same type of analysis for the more specific and a priori suspicious AE of rash. Note here the enhanced ability to detect drug related rash and also the virtually equal rates judged as unrelated for active and control patients. Finally, in Table 3, I present a similar analysis of data from a Canadian trial of the immuno-suppressant Cyclosporin A in rheumatoid arthritis (Tugwell, 1990). Here the relatedness judgement allowed a series of ordinal categories from "definitely related" to "definitely not related" and the reporting rate ratio for Cyclosporin vs placebo patients exactly follows the ordinal progression of relatedness with again the "probably not related" category having a value of about one.

This long-winded introduction to my discussion is essentially saying that many of the problems that Dr. Chuang-Stein is confronting in her paper could potentially have been "headed off at the pass" by changes in the reporting practices for AEs. By concentrating early on noise suppression in the reporting process, we can enhance our ability to statistically detect the true signal.

**Objective of Safety Analysis**

**a. Phase I and II**

The method proposed by Dr. Chuang-Stein for comparing the weighted intensities of adverse effects in a single multivariate comparison is ingenious and possesses great strengths. Being of the class of statistician that regards p-values as largely irrelevant in the context of safety, especially in Phase I/II trials, I tend to view this approach as statistical overkill. However, in certain situations, the technique may be very relevant and it does have the advantage of producing a relatively small set of summary statistics with which to characterize the negative consequences of one treatment over another.

At the Phase I/II stage of drug development, I believe the key objective in safety analysis is to tentatively describe and characterize the adverse effects of a new drug. In my experience true adverse reactions do not manifest themselves as a broad spectrum but as a few relatively specific types of experiences, some of which might be expected from what is known biochemically about the action of the drug, and others which are unexpected or even idiosyncratic. Rapid recognition of extremely troublesome side-effects may allow the early termination of a development plan but in general the weighting of therapeutic risks and benefits must await Phase III data. How one recognizes the relatively small set of characteristic toxicities in Phase I/II remains a problem in terms of the obvious scope for multiple comparisons and thus the increased likelihood of seeing unusually large differences in incidence by chance alone. The trick is to be able to group textural descriptions of similar events into a common set (not necessarily always from within a single body system), a tasks which is aided by the hierarchical structure of the WHO and COSTART coding systems. Ultimately, the recognition of candidates for the label of "side-effect" is a blend of statistical and biological/clinical judgement.

**Table 2: Rash in CATS**

Treatment Group	AEs	Relationship to Study Medication		
		Probably Not Related	Unknown	Probably Related
<b>PLACEBO</b>				
	Number Reports	39	16	7
	Rate /100 Assess.	1.25	0.51	0.22
<b>TICLOPIDINE</b>				
	Number Reports	33	38	42
	Rate/100 Assess.	1.19	1.37	1.51
Rate T/Rate P		1.0	2.7	6.8

**Table 3: All Adverse Experiences - CYRA**

Treatment Group	AEs Not Related	Relationship to Study Medication				Definitely Related
		Definitely Not Related	Probably Not Related	Possibly Related	Probably Related	
<b>CYCLOSPORINE</b>						
Number Reports	23	196	302	215	78	
Rate /100 Assess.	.8	32.0	49.3	35.1	12.7	
<b>PLACEBO</b>						
Number Reports	67	175	211	31	12	
Rate/100 Assess.	11.2	1.1	1.4	6.8	6.4	
Rate T/Rate P	0.3	1.1	1.4	6.8	6.4	

**b. Phase III**

The output of the controlled fishing expedition of Phase III safety data analysis is hopefully a relatively small set of characteristic drug reactions associated with an investigational drug that can target the safety review of Phase III. General reviews of adverse experiences are also required but the potential, if desired, for formal hypothesis testing, is enhanced by having only a few specific questions to address.

As might be anticipated by my earlier comments and my general tendency away from hypothesis testing for safety data, I view the eventual tradeoff of therapeutic benefit and risk as essentially subjective. The basic problem is not having a common quantitative unit with which to bring together benefit and risk. Our economist colleagues might argue that the measurement of patient utility offers a solution but in my view this is more theoretical than practical. I am firmly of the belief that in general one should describe the benefits and risks separately in their own natural units and let people decide if the trade off is reasonable. Initially, this will be the FDA, later on individual physicians and/or patients, and finally hospitals and third party insurers. It is quite possible that these decision makers will reach different conclusions about the relative merits of competing agents, as is happening currently with the various thrombolytics available. Clearly in some situations total mortality is the only relevant outcome and naturally combines at least part of the benefit/risk tradeoff. Occasionally, a single major adverse experience may be directly combinable with the primary efficacy outcome to reflect the net position. Examples of this situation might be ischemic stroke plus intra-cerebral hemorrhage in the evaluation of anticoagulants in patients with chronic atrial fibrillation or DVT plus major bleeds in post surgical patients treated with prophylactic anticoagulation.

**c. Post Marketing Surveillance**

To be effective, I believe post marketing surveillance of drug safety has to be structured, mirroring in part the way AE information is collected in clinical trials, although by definition unblinded and usually uncontrolled. In other words, individual practices have to be recruited to act as sentinels to reliably report all AEs associated with a particular product. Unfortunately, in my experience, the majority of post marketing surveillance is conducted via a much more ad hoc process and thus produces poorer quality data. I must admit my somewhat jaundiced view of the reliability of post marketing surveillance is colored by an earlier experience with the drug Suloctidil. This platelet active agent had been marketed in Europe for a number of years prior to us conducting a North American based trial (Gent, 1985) in stroke patients. It did not take us long to become aware of Suloctidil's hepatotoxicity and, in the face of only a modest trend towards efficacy at the interim analysis, decided to stop the trial early. The lack of appreciation of the

problem with liver toxicity prior to our study was even more surprising when, following the publication of our findings, a re-evaluation of available post marketing data in Europe did reveal a clear indication of the problem. This reevaluation lead quite rapidly to the worldwide withdrawal of the drug. While this anecdotal experience may have less relevance in North America, it does support, what I think we all accept, that statistical analysis of safety data is much easier if we know what we are looking for!

**Integrated Safety Summaries**

As an academically-based biostatistician with experience in clinical trials, I rarely get involved in formally integrating data over studies and then only on the efficacy side of the equation. The various forms of meta

analysis have gained great popularity following Peto's (1980) attempts to extract reason out of the apparent chaos of the early aspirin post-MI trials. Although largely accepted as the methodology of choice, some conceptual problems linger in terms of whether, and if so how, to incorporate between study variance into the computation of the pooled estimate of efficacy and its associated confidence interval. In theory, the same techniques are applicable to the safety side of the equation but I suspect that there is more scope for between study variation simply because of the open-endedness of the reporting process compared to that of the efficacy outcome.

**Use of Reference Ranges**

A case can be made that the utilization of reference ranges, like some forms of religion, does more harm than good! In general, I prefer to summarize laboratory data quantitatively rather than qualitatively. One might argue that this places even more premium on inter-laboratory standardization (unlikely to be attainable) or the use of a central laboratory. However, in practice, the use of change from baseline provides a measure of protection against inter-laboratory shifts in mean. An analysis of change for baseline divided by say one quarter of the reference range would, in addition, allow for inter-laboratory differences in variance.

If, for whatever reason, one must go ahead with a qualitative analysis based on reference range, then the litany of problems discussed by Dr. Chuang-Stein inevitably ensues. The fact that the patients we recruit often have active disease processes underway makes it unlikely that the proportion of subjects falling outside the range has any real meaning in itself. This does not, of course, invalidate a comparison between treatment groups in terms of the proportion (dare I say) "abnormal" nor does it imply that some process to define comparable reference ranges for each lab is a complete waste of time. The proposal by Oliver and Chuang-Stein to define these ranges within a study based on pre-treatment data plus comparative inter-laboratory assay of aliquot samples makes eminent good sense and it will be interesting to see if this methodology is implemented by industry.

**Concluding Comments**

There is no doubt in my mind that the open-ended nature of safety data creates an inherent analysis challenge in individual studies let alone in the bringing together of many studies into an integrated synthesis. Dr. Chuang-Stein's paper is an honest appraisal of the difficulties which points to possible solutions while not glossing over the problems. As such, it is a valuable source of guidance to all of us struggling to create order out of semi chaos. I was going to phrase the concluding remark in terms of making "silk purses out of sow's ears" but this is too pessimistic a note on which to leave this important issue!

## References

- Gent M, Blakeley JA, Hachinski V, Roberts RS, et al. (1985): A Secondary Prevention, Randomized Trial of Suloctidil in Patients with a Recent History of Thromboembolic Stroke, : 16 (3):416-424.
- Gent M, Blakely JA, Easton JD, Ellis DJ, Hachinski VC, Harbison JW, Panak E, Roberts RS, Sicurella J, Turpie AGG and the CATS Group (1989): The Canadian American Ticlopidine Study (CATS) in Thromboembolic Stroke - Results of a North American Trial. *Lancet*, 1:1215-1220.
- Peto R (1980) Editorial. *Lancet* 1:1172-1173.
- Tugwell P, Bombardier C, Gent M, Bennett KJ, Bensen WG, Carette S, Chalmers A, Esdaile JM, Klinkhoff AV, Draag G, Ludwin D, Roberts RS (1990): Low-dose Cyclosporin versus Placebo in Patients with Rheumatoid Arthritis. *Lancet* 335:1051-1055.

## Discussion

### David Salsburg

*Pfizer Central Research, Pfizer, Inc.*

Dr. Chuang-Stein has presented a sweeping examination of the statistical aspects of safety analyses for a new drug. In such a broad presentation, it is impossible to get into great detail about methods or presentations. I recommend that readers of this discussion go back and read (or reread) the Chuang-Stein, Mohberg, and Musselman paper in *Statistics in Medicine*. In this paper, the authors conquer many of the problems in safety analysis with a simple device. They require that the medical professionals examine the data and categorize each patient in terms of severity of adverse events by body system. As a result, the analysis deals with a well defined medically important set of events.

Without such a reduction of the data, what we call "safety data" consists of a complex mixture of records and observations, many of which are only indirectly related to the questions at hand. This mixture of information extends across the time (since patients are followed for periods up to three years), across dose (which may change over time), across types of observations (frank side effects, signs of disease, blood chemistries, etc.), across body systems or syndromes, and across patient types (male/female, aged/young, black/white, etc). It is impossible to display events across all five of these dimensions, and so we take slices across body systems, across dose, with time involved or with time ignored. The problems faced in analysis are how to present a small number of useful summary slices and how to determine "if anything happened."

There are some things that make no sense. As Dr. Chuang-Stein points out, it makes no sense to equate statistically significant changes in blood chemistries with lack of safety or the lack of statistical significance with safety. Most drugs are xenobiotic substances, and the ordinary metabolism of these drugs will cause changes in some of the blood chemistries. With enough patients, these changes reach formal significance. Most times, the changes are anomalous, such as increase in one liver enzyme and decreases in another. Dr. Chuang-Stein points out that the "normal ranges" are fictions that do not imply a normal distribution across patients and are often based on lab replications in addition to standard sets of patients. However, the "normal range" does provide a starting point when using blood chemistries as indicators of possible drug toxicity. It has been my experience that drug toxicity is associated with dramatic changes in blood chemistries. When drugs cause leukopenia, both the white blood count and the leukocytes drop well below their normal range. Drugs that involve liver toxicity cause SGOT to rise 10 to 100-fold.

Contrast this with the NSAID drugs, all of which cause a slight rise in BUN, that goes back down when the drug is removed, or the beta-blockers which cause a slight rise in cholesterol. There may be long term consequences of these changes, but these cannot be evaluated in the context of a new drug development program.

This brings me to my final point. There is no way of knowing of a new drug is "safe". In a drug development program, we try to anticipate problems that might arise when the drug goes into general use in the face of imperfect knowledge. It is a good idea to think of the potential danger of a new drug in terms of 2x2 table where the rows represent

- Events that might be expected from the drug's pharmacology
- Events that are totally unexpected

and the columns represent

- Events with a high enough frequency to be observed in a development program
- Events that are "rare" or idiosyncratic.

Examples of high frequency events that might be expected from the drug's pharmacology are pedal edema with a calcium channel blocker, or postural hypotension with an alpha-blocker. These events are usually dose related, and the safety question becomes one of the frequency, severity, and toleration for the expected range of therapeutic doses. An example of a rare event that might be expected is the interaction between MAO inhibitors and the ability of a patient with genetic defects to digest certain amino acids.

Unless the drug is a unique agent for a deadly or highly morbid disease, the occurrence of an unexpected event of sufficiently high frequency means that further development of the drug will be killed, if not by the company, then by the regulators. The rare unexpected event is the bugaboo of all regulators. There is no way we can identify the event in a 2000-3000 patient developmental program. And, even a careful phase IV trial will often fail to locate such events, or they will occur so rarely as to cast doubt on whether they are drug-related. Unless society wants to stifle all future drug development, society must learn to live with that uncertainty.

Thus, safety analysis consists of what can be done with the three cells in the X table. The expected high frequency event can be characterized through the use of logistic regressions and estimated hazard functions. The rare expected event can be sought by studying peculiar and potentially susceptible patient populations or it can be mentioned in the package insert to warn such patients away from the drug. The statistical program addressed by much of Dr. Chuang-Stein's paper deals with techniques for locating the relatively high frequency unexpected event. I do not have any better answers than she does.

## Discussion

### R. Srinivasan

*Division of Biometrics*

*The Food and Drug Administration*

The paper by Christy Chuang-Stein has addressed important and frequently arising statistical issues concerning the description and evaluation of safety analyses information in clinical trials. A rational position to take in the development of new drugs is to design trials to provide definitive information about efficacy rather than safety. Safety issues have been monitored in each trial in an effort to describe the drug safety profile. Furthermore, such data is accumulated across trials in an effort to develop a more comprehensive safety profile. Some statisticians employ inferential methods to describe safety data and others use descriptive

methods and p-values obtained from parametric methods for exploratory purposes. Since the safety data are multi-faceted, the author's proposal to combine all relevant safety data by body systems and analyze the accumulated information using a multivariate approach is interesting. However, I usually evaluate the total number of adverse clinical experiences for a given body system, say Central Nervous System (CNS). If it is significant, I then look at individual components because these are what need to be noted in the Label for physician's use.

Another important issue in the analysis of adverse events is multiplicity. The number of statistical tests will increase with the increase in the number of events, among which some will have low p-values (e.g.,  $p \leq 0.10$  two sided) and others will have higher p-values. Enas (1991) and Huster (1991) have also noted that a higher p-value (e.g.,  $p \leq 0.25$  two-sided) does not necessarily imply lack of association between treatment and an adverse event because of relatively low power that is available to detect such association with the available sample size in most clinical trials (or their meta-analysis combinations). According to Koch (1991), for any study, an important role of analyses of adverse events is to identify suggestive trends for which more focused evaluation is applied subsequent to integration with other studies.

The author, along with Mohberg and Sinkhole (1991), has proposed three ways to incorporate benefits and risks into one analysis. The first extends Hilden's (1987) procedure while the others generalize the benefit/risk ratio considered in Paine and Lokan (1975). All three procedures employ weights to compute the summary statistics. The author's concept of benefit-risk assessment to be made with respect to the underlying disease or symptom is also reflected in Japanese Protocols as a "usefulness" criterion to be assigned by the investigator. The author recommends conducting additional benefit/risk assessments in the presence of newly detected side effects after a new drug/treatment has entered the market. In general, this is not practical to accomplish in the regulatory milieu.

The author advocates that as in efficacy evaluation, meta-analysis can be carried out to pool safety data from different studies. She points out that when pooling data from different studies, one has to pay attention to the nature of the problem, dose received and the exposure duration. The author advocates a Bayesian approach for the parameters of interest, which incorporates a variance component that takes into account between-study variability. Statistical evaluation procedures such as logistic regression with covariates may be used to adjust imbalances due to age, gender, race, underlying disease status etc., and help explain observed problems.

Laboratory results provide very useful information regarding a patient's safety in a trial. Lab-specific ranges are commonly used in multicenter trials where patient's specimens are processed at the respective study sites. The use of lab-specific reference ranges which are based on normal population values misses two aspects: (1) a shift to high or low, but not out of "normal range", and (2) disease x adverse laboratory experience x adverse clinical experience interaction. The author strongly feels that safety monitoring should focus on assays for assuring the continuation of a pre-existing status of a patient's biochemistry. It is a fact that statisticians are not aware of a universal definition of reference populations and a common method to establish reference ranges. The author feels that these two deficiencies along with the differences in the assay procedures and the accuracy in diagnosis raise doubts about the usefulness of reference ranges in multicenter studies.

The author recommends that each laboratory establish a laboratory specific quality control proficiency baseline based upon analytical results from a standard, common set of specimens. Each patient's pre-treatment laboratory results, adjusted to the proficiency baseline will then serve to identify abnormal laboratory results. Such adjustments help to

consolidate data from different laboratories within one study where differences occur as a result of laboratories using different procedures or equipment rather than results. The author feels that the current practice of constructing reference ranges based on data from normal populations is not appropriate when laboratory results of body systems are affected by drug or disease. I would think that one measure of therapy success is for function and values to return to "normal."

I wish to thank Dr. Robert T. O'Neill, Director, Division of Biometrics, Food and Drug Administration, for providing me the opportunity to be a discussant, Dr. Ralph Harkins, my supervisor and my colleague, and Ms. Beth Turney for offering useful suggestions in preparing this discussion.

## References

- Chuang-Stein, Mohberg and Sinkhole (1991). Three measures for simultaneously evaluating benefits and risks using categorical data from clinical trials: *Statistics in Medicine*, 10:1349-1359.
- Enas GG (1991). Making decisions about safety in clinical trials - the case of inferential statistics. *Drug Information Journal*, 25:439-446.
- Hilden J (1987). Reporting clinical trials from the viewpoint of a patient's choice of treatment. *Statistics in Medicine*, 6:745-752.
- Huster WG (1991). Clinical trial adverse events: The case of descriptive techniques. *Drug Information Journal*, 25:441-456.
- Koch GG. (1991). Discussion: Statistical Perspective. *Drug Information Journal*, 25:461-464.
- Paine JT and Likens MK (1975). A survey of the benefits and the risks in practice of Radiology, *CRC Clinical Reviews in Clinical Radiology and Nuclear Medicine*, 6:425-429.

## Discussion

Janet Wittes

*Statistics Collaborative*  
Washington, DC 20036

My introduction to the analysis of safety data in clinical trials was during my first Data and Safety Monitoring Board Meeting. Confronted with long lists of laboratory parameters, symptoms, signs, and events, I did what any rational person would do—I called Max Halperin. How, I asked, was one supposed to make sense of all these numbers? How could anyone possibly decide whether a drug had an unacceptable safety profile? He laughed and gave me sage advice, "Tune out during the discussion." After ten years of tuning out, I welcome Dr. Chuang-Stein's thoughtful discussion of how to collect, analyze, and interpret safety data. She so ably points out that safety should not be viewed in isolation from efficacy; ultimately, we need to ask questions about the whole person. Are people treated with a specific drug better off, overall, than people treated some other way? As she says, "...an overall safety profile can be drawn based on data from all sources. This requires an effort to intelligently combine all safety information to facilitate the overall summarization." Her interesting previous work with Mohberg and Sinkula (1991) integrated safety and efficacy variables.

Her current paper discusses a number of important topics, among them, the role of p-values and multiple significance tests in evaluating the safety of a drug, methods of combining safety data from many trials, and the ambiguity of "reference values" or "normal ranges" in the context of clinical trials on people with disease. Lurking behind all the discussion is the problem of assigning medically meaningful levels of importance to the many

## Journal Response

The following letter is Reprinted from *CLINICAL PHARMACOLOGY AND THERAPEUTICS* (1992) 52:104-105, with permission. (Copyright 1992 by Mosby-Year Book, Inc.)

### Statistics and Clinical Trials

#### To the Editor:

The American Statistical Association is a 150-year-old professional society representing approximately 16,000 statisticians in North America and throughout the world. The Biopharmaceutical Section comprises about 1,000 members with primary interest in clinical trials and other areas of pharmaceutical research. We wish to comment on the recent article in the Journal by Dr. Lewis Sheiner<sup>1</sup> in which he made a very strong statement that many individuals in the clinical research community are led to simplistic objectives, designs, and analyses of their trials by their ignorance of statistical concepts or by statisticians who cannot match their theory with clinical reality.

---

**Clinical research is a multidisciplinary responsibility. The objective must be clearly conceived and presented, it must be clinically relevant, and it must be attainable as designed.**

---

Although we agree with much of what Dr. Sheiner presents in his article, we believe it lacks balance and the tenor is one that seems to minimize the importance of statistical considerations in clinical drug evaluation. In fact, his solution to the illness as he sees it is that "clinicians must regain control over clinical trials..."

Dr. Sheiner properly condemns the inappropriate use of such things as statistical methods, analyses of convenience, and inadequate conceptualization of objectives. However, these are not failings that can be broadly attributed to statisticians but must be shared by all principals in a trial.

In his first example, the use of an irrelevant analysis is properly criticized. However, the example cited does not include a single statistician in the list of authors.<sup>2</sup> If a properly qualified statistician were an integral member of this team, perhaps a more appropriate analysis might have been conducted.

Dr. Sheiner states in this example that "we are not encouraged to carefully model..." and "We are instead encouraged to focus on a dull and featureless null model..." We must question *who* is "not encouraging" and *who* is "encouraging" and, more important, why this is at issue. A good scientist does not let lack of encouragement by some unnamed third party (perhaps a statistician) compromise the quality of his or her research.

The "test the null hypothesis" mind set is, unfortunately, a pervasive disease that infects both clinicians and statisticians. But it is important to identify the mind set as the disease, not hypothesis testing. There are many important uses for these methods, but they are not appropriate unless there is a clinically important hypothesis to test. If such an approach is inappropriate, the team should not

use it. Inertia is not an excuse for use of methods that are convenient but that do not address the objectives of the research. Popularity is not the best method for determining statistical approaches in either the planning or evaluation of any scientific study. No professional statistician would ever condone routine, unthinking analysis.

Dr. Sheiner also expresses a concern regarding statisticians' restricting a clinician's ability to modify a protocol by looking at particularly accumulated data. There are methods for doing exactly this, and any trained statistician should be familiar with their use. They do not, and should not, allow unrestricted review and modification. Reviewing data in an open manner in clinical trials can create large and immeasurable bias. No scientist would want to review data to satisfy their curiosity if the scientific validity of the study would be compromised. All research involves risk, but to evaluate these risks properly, we must know the chance that we will be wrong and the magnitude of any bias that may exist. Without this knowledge, our decisions will be only guesses.

According to Dr. Sheiner, statisticians have done little to educate clinicians about basic statistical concepts. We must certainly accept responsibility for this shortcoming, and we are able and willing to do much to correct this situation. But to be successful, we need the commitment and support of medical schools in two areas. First, biostatistics and clinical trial method should be a required course, one that must be passed and taken seriously, if we are to change the ignorance that is too widespread. Second, physicians in training must be encouraged to appreciate the true value and need for statistical thinking (not statistical methods as is often taught) in clinical research. We, as statisticians, have an obligation to teach our profession, but if clinicians are to be able to use this body of knowledge in research, they must have a desire to learn.

The final section of Dr. Sheiner's article, "Reasserting Epistemologic Authority," is right on target with one exception, and that is the statement with which we began this letter: "clinicians must regain control over clinical trials..." Clinical trials require a multidisciplinary approach; it is as inappropriate for a clinician to direct a clinical study without appropriate statistical expertise as it is for a statistician to design a study without clinical participation.

The design, conduct, and evaluation of clinical research is not a uniquely clinical responsibility, and it is not one that falls on the shoulders of the statistician. Clinical research is a multidisciplinary responsibility. The objective must be clearly conceived and presented, it must be clinically relevant, and it must be attainable as designed. It is not just wasteful to perform poorly designed and conducted studies or to perform a routine analysis that does not address the real clinical objective of the study; it is unethical to expose any human beings to experimental therapy without a clear knowledge of the likelihood of success.

Clinicians and statisticians have a personal and collective responsibility to defend this trust by optimizing the use of our technical and intellectual skills and resources. We must share our expertise toward more effective and efficient clinical research by working zealously to improve and develop the understanding of each other's profession that is essential for excellence in clinical research.

**Bruce E. Rodda, PhD, Chair-elect**  
**Camilla Brooks, PhD, Chair**  
**Gladys Reynolds, PhD, Past Chair**  
 Biopharmaceutical Section  
 American Statistical Association

#### References

1. Sheiner LB. The intellectual health of clinical drug evaluation. *Clin Pharmacol Ther* 1991;50:4-9
2. McQuay HJ, Carroll D, Frankland T, Harvey M, Moore A. Bromfenac, acetaminophen, and placebo in orthopedic postoperative pain. *Clin Pharmacol Ther* 1990;47:760-6.

measures made in assessing safety. What is reasonable calculus for assigning weights to abnormal laboratory values and clinical events? When we deal with efficacy endpoints, my own response to selecting weights meaningful to physicians, patients, and families tends to favor subjective assessment of the entire complex of events an individual experiences (Follman, Wittes and Cutler, 1992). For adverse events and safety data, the combination of subclinical findings and clinical events is so complicated that I am pessimistic about finding generally applicable approaches that parsimoniously characterize experience.

Below I address two specific issues. Chuang-Stein alludes to the first topics, the problem of how to deal with the fact that adverse events and safety data contain many correlated variables. We are often told that clinical trials are too small to identify potential adverse effects of therapy. Sometimes, the numbing effect of long lists of variables masks true adverse effects. Rare adverse events are especially difficult to detect in clinical trials because an individual trial, or even a group of trials, may include too few people for an infrequent event to manifest itself. I believe, however, that underestimation of event rates is not the only problem, because the nature of the way we report safety data often inflates the apparent adverse experience on drug. An elderly person who falls and breaks her hip in a clinical trial may show up in tables under "falls," "fainting episodes," "dizziness," "hypotension," "osteoporosis," and "lightheadedness." Thus, a slightly higher number of falls in one treatment group may be magnified by the nature of our reporting. Although the preparer or reader of the safety report may be well aware that a single event can generate entries in many categories of safety variables, nonetheless the cumulative effect of seeing many categories with excess event rates often leads to an exaggerated sense of the toxicity of a study arm. Chuang-Stein's various suggestions concerning multivariate approaches are useful: my own preference is to categorize events and laboratory values not only by organ system, but by "syndromes" as well. Then, each cluster of events counts only once, so that a single adverse event is not counted several times. Whatever we do, we should adhere to her appeal that we think about systematic ways to collect data on safety.

The second issue is the ongoing monitoring of safety. Much of Chuang-Stein's paper addresses methods for sensibly integrating data on safety after a study is over as well as approaches for combining results from one or more clinical trials with results from available post-marketing studies. At that point, there is often available a relatively static set of data that includes reliable information on efficacy. By contrast, most of my own experience in dealing with safety data occurs not at the end of data collection, but rather during the monitoring of studies. Data and Safety Monitoring Committees usually have three basic responsibilities. The simplest of the tasks is the determination of how well the study is continuing administratively. The third task of the Committee is to evaluate efficacy. Often, the study includes guidelines for early termination that assist the Committee in its deliberations. Here I refer to the second task, determining during the course of the study if the new therapy is sufficiently harmful to recommend termination of the study or alternation of the dose or mode of administration. Many of the problems that Chuang-Stein discusses are even more acute in the process of data monitoring because the data on adverse events and safety are so sparse and because the information on efficacy, which would allow at least an informal balancing of benefit against risk, are almost unavailable. During monitoring, the question of p-values is often moot. If the adverse event is harmful enough, we may not want to continue a study long enough to prove harm. On the other hand, if the safety data show minor laboratory abnormalities in a serious disease, we may well judge that even a statistically significant result is not sufficiently worrisome to necessitate

changes. And sometimes, such as in trials of immunosuppression, early adverse events herald efficacy.

As Chuang-Stein observes, "Because of the diverse nature of safety data that are routinely collected in a trial, the techniques to summarize them are less obvious and straightforward compared to those of the efficacy data...." Precisely because of this complexity, physicians and statisticians must jointly design collection methods for safety data, select approaches to summarizing those data, and interpret the results.

## References

- Chuang-Stein C, Mohberg NR and Sinkula MS (1991). Three measures for simultaneously evaluating benefits and risks using categorical data from clinical trials. *Statistics in Medicine*, 10, 1349-1359.
- Follman D, Wittes J, and Cutler JA (1992). The use of subjective rankings in clinical trials with application to cardiovascular disease. *Statistics in Medicine*. 11, 427-437.

## Letter from the Editor

This issue of the *Biopharmaceutical Report* focuses on the important issue of safety analysis. We have an excellent paper by Christy Chuang-Stein and discussions from industry, academia, and government perspectives. Even with these well-thought out discussions, my feeling is that this is only the tip of the iceberg on this issue.

We have recently done a brief and not quite random survey of Section members and realized that some of you did not receive one or both of the first issues of the *Biopharmaceutical Report*. If you received this issue, that means that you are a member of the Biopharmaceutical Section. If you wish to receive the previous issue, the Membership Department of ASA has a limited number of copies it can still send out. Please write or call them to request issues 1 or 2 at:

1429 Duke Street  
Alexandria, VA 22314-3402  
Tel. (703) 684-1221  
FAX: (703) 684-2036

**Avital Cnaan**  
Editor

### Editorial Board

**Avital Cnaan**  
University of Pennsylvania  
Editor

**Thomas Bradstreet**  
Merck Research Laboratories  
Associate Editor

### Layout and Design

**Alison Stern-Dunyak**  
American Statistical Association

## Book Review

***Design and Analysis of Bioavailability and Bioequivalence Studies.* Shein-Chung Chow and Jen-Pei Liu. New York: Marcel Dekker, Inc., 1992. x + 416 pp.**

**Reviewed by: Carl M. Metzler**

*The Upjohn Company*

Stimulated by the dual pressures for quality pharmaceuticals and lower costs of those drugs, the topic of bioavailability and bioequivalence has been discussed often in the last 20 years. The continuing frequent international conferences devoted to this topic suggest that not all the issues have been resolved. This book is an extensive and exhaustive coverage of the statistical methods developed to support the pharmaceutical and medical sciences. The references are generally broad and complete through 1991, although the many citations to unpublished work of the authors is not helpful.

As befits its presence in Dekker's statistics series, this book is better suited to statisticians than to other scientists. The mathematics is extensive and complete. For example, in an interesting chapter on evaluating bioequivalence with clinical endpoints there is an extensive discussion of analysis of binary responses.

The book gives enough of the history and science of bioequivalence to make the book readable for those statisticians not familiar with the subject. As the authors indicate, most decisions about the bioequivalence of two formulations have been based on the assumption that the parameters that mark the bioavailability of each formulation have the same distribution except for a shift of location. Rightly then, most of the book is devoted to evaluating whether the shift in location is small enough that the two formulations may be judged as bioequivalent. There are two chapters on design, including a chapter on alternatives to the standard  $x$  crossover design. Most of the current topics of interest - transformations, differences in variances, outliers, individual bioequivalence - are covered. The only current topic not mentioned is bioequivalence of controlled release formulations.

Although judging this to be an outstanding book on the statistics of bioequivalence, this reviewer does have some criticisms. The authors seem to be limited by classical statistical concepts. Thus, they insist on putting the evaluation of bioequivalence into the hypothesis testing framework. This leads to some awkwardness or incorrect statements when discussing the use of confidence intervals to decide bioequivalence. For example, on pages 75 and 123 the authors are critical of the confidence interval approach because for values of relative bioavailability within the accepted interval the probability of declaring bioequivalence is not the level of the confidence interval. Although testing a null point hypothesis has little application in bioequivalence, the authors only discuss AOV for identifying and estimating variances after a lengthy discussion of such tests.

The book differs from most discussions of bioequivalence by naming as "carryover" effects which most other authors call "sequence effects". The book should at least relate the two. On page 18 and other places the authors repeat the common mistake of justifying the log-transform by the skewness of the observed data (eg, AUC). It is the distribution of the model residuals that is important. On page 143 the authors repeat a common criticism of the Anderson-Hauck procedure, neglecting to point out that since bioequivalence data are nonnegative they are never going to have a variance large enough for the criticism to be valid. The authors

could have provided more guidance for the non-statistician through the extensive mathematics. This and other criticisms may be more a matter of style than substance. For another survey of the subject matter of this book see a recent special issue of the *International Journal of Clinical Pharmacology, Therapy and Toxicology* [1].

### Reference

V. W. Steinijans and H.-U. Schulz (eds), Bioequivalence Assessment: Methods and Applications; *Int J Clin Pharmacol Ther Toxicol* 30: Supp. 1, (1992).

## Meeting Overview

### Sessions at the Joint Statistical Meetings sponsored by the Biopharmaceutical Section in Boston, August 1992.

Eleven sessions were sponsored by the Biopharmaceutical Section. Three of the sessions are summarized below.

#### ■ Longitudinal Analysis and Repeated Measures

**George W. Divine**

*Henry Ford Health Systems*

Most of the papers in this session offered solutions to problems presented by missing data in longitudinal studies.

The first presenter was HL Patel of Berlex Laboratories, Inc., who discussed "A Repeated Measures Design with Repeated Randomization." He detailed how such a study could be analyzed and pointed out that it could allow greater economy in recruitment time than more conventional designs. The University of North Carolina was well represented, as the source of three papers. The first of the UNC presentations was given by Ronald Helms, who was also a co-author of the other two UNC papers. His talk, entitled "Intentionally Incomplete Repeated Measures Designs for Clinical Trials," described how power for incomplete repeated measures designs might be estimated and showed that such designs can give increased efficiency. The second UNC paper was given by James Grady, who discussed "Modeling the Covariance Matrix for Incomplete Longitudinal Data." He presented graphical representations of the fit of various convenient structural covariance models. Mark Von Tress of Alcon Labs, Inc. made the fourth presentation: "Longitudinal Models for Polytomous Responses." He discussed how such data could be analyzed and also commented on some remaining technical issues. The fifth presentation was given by Ann Marie Kelly, of the LSU Medical Center. Her paper: "The Analysis of Longitudinal Data When a Portion of the Subjects Fail to Respond to Treatment," dealt with the impact of the mixture of distributions that will result in the indicated circumstance, and simulations were presented that suggest a fractional degree of freedom  $X^2$  distribution may give a good approximation for the likelihood ratio test statistic. Finally, Sandra Stagnate, the third UNC presenter, discussed "Multicollinearity in Mixed Models." She described a formal approach to the problem and described how diagnostics could be used in this situation.

In short, the session gave some interesting results concerning several difficult questions in longitudinal analysis.

## ■ Comparison of Approaches to Population Pharmacokinetic Modeling: A Case Study Using Clinical Data

Denise J. Roe

Arizona Cancer Center

Population pharmacokinetic models allow the estimation of pharmacokinetic parameters using routine clinical data. The goal of this session was to compare these techniques using an example data set. The data set contained serum quinidine measurements for 136 male hospitalized patients with atrial fibrillation or ventricular arrhythmias treated with oral quinidine. Eighty percent of the patients had one, two or three quinidine concentration measurements. The speakers were first asked to estimate the population pharmacokinetic parameters for a one-compartment open model with first-order absorption and first-order elimination, in terms of  $k_a$  (absorption rate constant),  $V$  (volume of distribution), and  $Cl$  (clearance). This initial analysis ignored the potential impact of demographic and clinical covariates, and was included to provide a simple comparison of the estimates. The speakers were then asked the more relevant clinical question of whether age, height, weight, race, smoking status, ethanol abuse, congestive heart failure, creatinine clearance, dialysis treatment, or alpha-1-acid glycoprotein concentration had an impact on  $Cl$ . Space constraints preclude a comparison of the estimated clearance after adjusting for appropriate covariates.

The first speaker was Stuart Beal from the University of California, San Francisco. He discussed the nonlinear mixed effects model, as implemented in the program NONMEM. The parameter estimates are shown in the table. Analyses of the effects of the covariates showed that the alpha-1-acid glycoprotein concentration had a large impact on clearance. Current ethanol abuse had a considerable influence on clearance. Height, weight, race, creatine clearance, and dialysis treatment each has a minor influence. NONMEM is well tested and documented, and has been used for multiple data analyses. The second speaker was Jon Wakefield from Imperial College. He discussed the Bayesian approach via Gibbs sampling. The mean and standard deviation of the distributions of  $Cl$ ,  $V$  and  $k_a$  are shown in Table 1. Clearance was related to alpha-1-acid glycoprotein concentration, creatinine clearance and weight. Graphical methods to detect outlying subjects were also mentioned. An exportable computer program is under development. The third speaker was Alain Mallet from INSERM. He discussed a nonparametric maximum likelihood approach. The mean of the distributions of  $Cl$ ,  $V$  and  $k_a$  are shown in the table. Clearance was related to alpha-1-acid glycoprotein concentration and history of ethanol abuse. A computer program

is available. Marie Davidian from North Carolina State University briefly presented a smooth nonparametric maximum likelihood approach, which will soon be available in the pharmacokinetic literature. Clearance was related to alpha-1-acid glycoprotein concentration, creatinine clearance and weight.

The discussant was Thomas Ludden from the FDA. He was involved in the original analysis of this data set using NONMEM, and summarized these results. His recommendation was that additional comparisons of data sets are needed, particularly data sets constructed to provide potential problems for the methods. This approach is being pursued by the Biopharmaceutical Section work group on Population Pharmacokinetic Modeling.

## ■ Issues in Dose Response and Drug Combination Studies

David M. Lansky, Ph.D.

Searle

The opening and closing talks discussed properties of and improvements upon the Cochran-Armitage trend test (CA test), respectively. The last talk was a selective tour of results from large simulations that was undertaken to evaluate the effect of dose spacing and various methods of combining control groups on the type I error rate. For those who use the CA test regularly, the written report should be quite useful. The first talk compared two versions of a modified CA test (the modifications allow fitting of a covariate), Peto's prevalence method, and logistic regression on several data sets. While the  $p$ -values from the four methods are generally similar, the modified CA test methods appear to give large  $p$ -values slightly more frequently. If the modified methods are in fact less powerful, the loss of power may be acceptable since the calculations are substantially simpler.

The second talk discussed the use of the four parameter logistic model with non-constant variance for assay and calibration. The authors draw a useful distinction between minimum detectable concentration (MDC) and reliable detection limit (RDL); explaining why RDL is usually larger than MDC. Unfortunately, they did not discuss the effect of sample size on RDL, MDC or the limit of quantitation. The calibration intervals are markedly affected by the choice of the variance function. While inversion is perhaps the preferred method of interval construction, this can produce infinite intervals, while Wald intervals are symmetric and (usually) bounded. The third paper discussed a semi-parametric compartment model for development of two tumors. The model allows transitions from each of four states (tumor-free, tumor A only, tumor B only, and tumors A and B) to death and transitions to states with more tumors. This model is particularly appropriate for small data sets where there are two competing causes of mortality.

The fourth presentation was a good complement to the morning session on "Statistical Methods for the Detection of Interactions Between Drugs". Here we saw that by defining additivity between drugs to mean statistically independent action we gain a scale-invariance property. The major point of the presentation was that the application of Lemke's algorithm for smoothing the density estimate could make local drug interactions apparent. While work is needed to characterize the properties of this smoothing method in this setting, the approach appears promising. After the presentation, a questioner was concerned that this definition of additivity could imply that a drug is not additive with itself.

Table 1. Parameter Estimates for Oral Quinidine (Ignoring Covariates)

	Cl Est.(C.V.)	V Mean (s.d.)	$k_a$ Mean (C.V.)
Nonlinear Mixed Effects Model	11.6 (32%)	251.0 ( 8%)	1.15 (40%)
Bayesian Analysis via Gibbs Sampling	12.4 (.46)	271.4 (17.7)	1.65 (.76)
Nonparametric Maximum Likelihood	12.4 (32%)	225.4 ( 6%)	0.94 (33%)

# Software Review

## A Taste of S-Plus

**A. Lawrence Gould**  
Merck Research Laboratories

The S-Plus language, widely used in academic and some industrial settings, appears to remain largely lingua incognita for (at least) non-academic biopharmaceutical statisticians. However, S-Plus can be quite valuable for certain applications, and even may provide unique capabilities. The system can be run under DOS as well as UNIX, so it is accessible to users with DOS-compatible PCs. Rather than sing the praises of S-Plus, which the vendor of S-Plus can do better than I can, I would like to illustrate the utility of S-Plus with a couple of applications that arose recently. These by no means approach, let alone stretch, the limits of S-Plus, but they do provide some flavor of what can be accomplished using S-Plus. S-Plus is available from Statistical Sciences, Inc., 1700 Westlake Avenue North, Suite 500, Seattle, WA 98109; (206)-283-8802.

**Example 1:** Relating visual field threshold sensitivity (expressed as the logarithm of the threshold sensitivity measurements averaged over the values obtained for a number of locations on the visual field using an automated perimetry device) to age and a measure of lens density. Each subject provides three measurements: TS (threshold sensitivity), LD (lens density) and Age. The data started out as a simple ASCII file like this:

```
26 0.52 2.150      100 rows like this
21 -0.18 2.475
71 0.87 1.762
72 0.67 1.650
59 0.62 1.875
```

S-Plus allows data arrays to be read into the system very much like SAS datasets. Before doing this, it is handy to add the names to your data file, like so:

Age	LD	LogMean	Insert this line
26	0.52	2.150	Dataset now has
21	-0.18	2.475	101 rows
71	0.87	1.762	
72	0.67	1.650	
59	0.62	1.875	

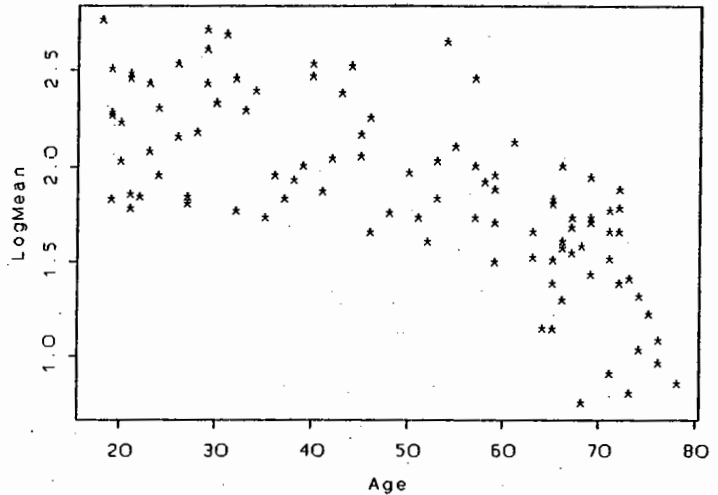
Suppose these data are in the DOS file "g:visflds.dta". They can be brought into S-Plus as what S-Plus calls a **table frame** (analogous in many respects to a SAS dataset) by

```
visflds <- read.table("g:visflds.dta",header=T)
```

The variables can be referred to hereafter by the names given in the first line of the input dataset. To save having to refer explicitly to visflds, execute the command

```
attach(visflds)
```

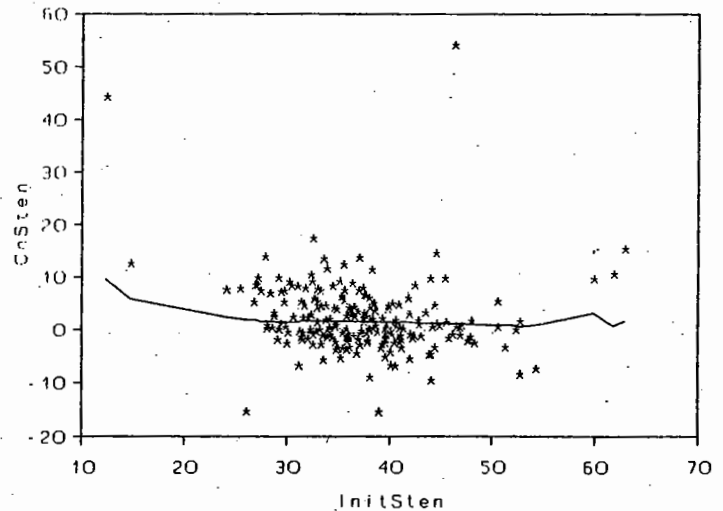
Now let's look at a plot of the data: `plot(Age,LogMean)` yields



adding a smooth curve using a kernel smoother with bandwidth 15:

```
visflds.ks <- ksmooth(Age,LogMean,bandwidth=15)
```

creates the smoothed values; add them to the plot with `lines(visflds.ks)` to get



A quadratic or cubic in Age might fit LogMean well. Similar steps suggest that LogMean is linearly related to LD. At worst, therefore, LogMean might be related to Age and LD by

$$\text{LogMean} \sim \text{Age} + \text{Age}^2 + \text{Age}^3 + \text{LD} + \text{LD}^2 + \text{Age} * \text{LD}$$

in S-Plus's notation for expressing models. A simpler model might do. There are only 6 predictors, so it is practical to consider all 64 possible models.

We can keep the predictor value ranges from getting too wide by standardizing them and adding the new variables to the table frame:

```
visflds$stage <- (Age - mean(Age))/(max(Age) - min(Age))
visflds$std <- (LD - mean(LD))/(max(LD) - min(LD))
```

Note the way S-Plus refers to variables within a table frame; every language has its idiomatic expressions, and S-Plus is no exception. Now add some more variables to visflds

```
visflds$stage2 <- tage^2;      visflds$stage3 <- tage^3
visflds$tld2 <- tld^2;       visflds$tageld <- tage*tld
```

A list of the variable names would be convenient here:

```
vnms <- c("tage", "tage2", "tage3", "tld", "tld2", "tageld")
```

Now for the fun part. First, construct a matrix of explanatory variables from the table frame (note how easily this is done):

```
xvars <- visflds[vnms]
```

Next, and this is a very powerful capability, find all of the regressions:

```
allregs <- leaps(xvars, visflds$LogMean, labels=vnms)
```

Printing the results takes a bit of work, but actually is not much worse than using PUT files in SAS:

```
attach(allregs)      Construct explicit output matrix
allregs.out <- cbind(size,Cp,labels)
sink(file="g:allregs.out") Direct output to DOS file
allregs.out          Write the output matrix (to file)
sink()               Direct output back to terminal
```

Here are the key results from the exploration of all possible regressions: (Good models have  $C_p \leq p$ )

p	C <sub>p</sub>	predictors	
2	11.33	tld	
2	45.72	tage	
3	7.09	tage, tld	
3	10.09	tage3, tld	
3	11.59	tage2, tld	
4	2.57	tage, tage2, tld	"Best" model
4	5.29	tage2, tage3, tld	
4	7.87	tage, tld, tageld	
5	4.08	tage, tage2, tld, tageld	
5	4.45	tage, tage2, tage3, tld	
5	4.46	tage, tage2, tld, tld2	
6	5.43	tage, tage2, tld, tld2, tageld	
6	5.80	tage, tage2, tage3, tld, tageld	
6	6.26	tage2, tage3, tld, tld2, tageld	
6	6.29	tage, tage2, tage3, tld, tld2	

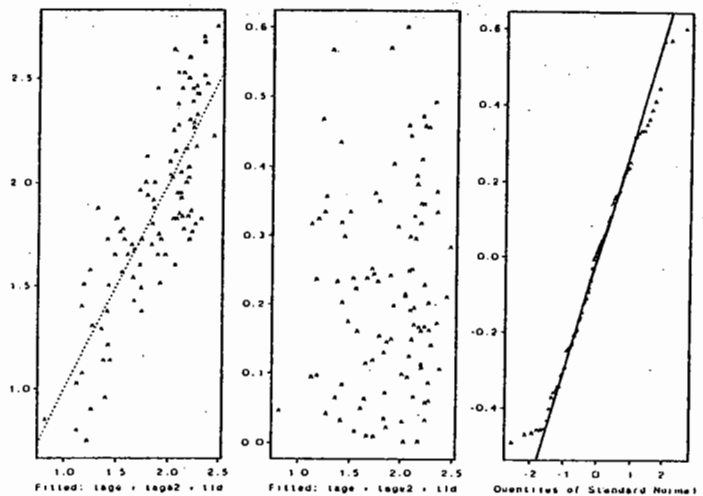
This simple, easily executed analysis suggests a sensible, intuitively attractive model.

To see if the model actually fits the current dataset well, fit it

```
visflds.lm <- lm(LogMean ~ tage + tage2 + tld, visflds)
```

and construct diagnostic plots

```
par(mfrow = c(1,3))      Arrange plots on page
plot(visflds.lm)         Object-oriented magic
qqnorm(residuals(visflds.lm)); qqline(residuals(visflds.lm));
```



Ordinate of first plot is LogMean, ordinates of 2nd & 3rd plots are residuals

The first diagnostic plot suggests a reasonable fit. The second plot reveals no pattern in the residuals as a function of the fitted values. The third plot suggests the residuals are nearly normally distributed.

**Example 2:** Robust weighted general linear models with heavy-tailed data from an atheroma study

Kind of data supplied (with variable names added)

Alloc	Clinic	Trt	nLesion	InitSten	FinSten	ChSten
1	B	A	11	35.35	42.27	7.92
3	B	A	2	40.56	40.90	0.34
5	B	A	7	34.08	35.77	1.69
-						
1504	D	A	5	26.71	31.58	4.87
1505	D	A	12	27.07	36.57	9.50
1518	D	B	8	29.35	30.13	0.78

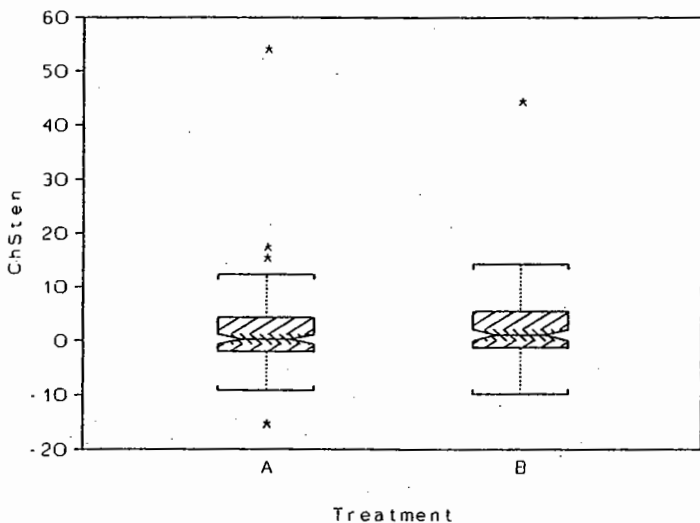
These are read into S-Plus just as in Example 1. "Clinic" and "Trt" have non-numeric values; S-Plus will treat these as factors, essentially the same as SAS CLASS variables. "nLesion" is the number of lesions, "InitSten" is the average percent stenosis (closure) of the lesions as measured by coronary angiography initially, "FinSten" is the same at the end of the study, and "ChSten" is the arithmetic difference between the initial and final scores.

The relationship between ChSten and Trt, InitSten, and Clinic is what we are after.

First step, as before: plot data — this time use boxplots to see if there will be an outlier problem

```
attach(qcasall) qcasall = table frame with data
boxplot(split(ChSten,Trt),yaxs="e",notch=T,outchar=T,outline=F)
```

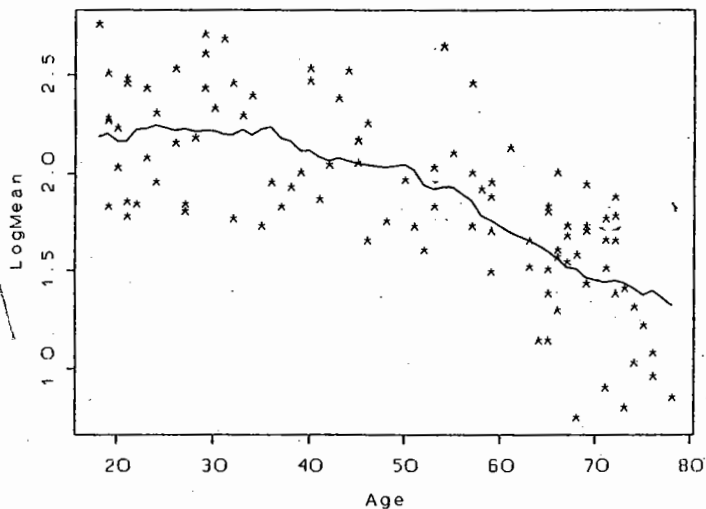
yields



The distribution seems to have heavy tails: the lines go out to 1.5 times the interquartile range. Also, the observations have differing precisions, as indicated by the values of nLesion. Thus, we have data with varying precision and a heavy-tailed distribution. Ugly, but common.

What sort of relationship appears to exist between ChSten and InitSten? Hit it with the smoother.

```
qcasall.ks <- ksmooth(InitSten,ChSten,bandwidth=30)
plot(InitSten,ChSten,xaxs="e",yaxs="e")
lines(qcasall.ks)
to get
```



Looks like a quadratic at most — also, there are outlying values for InitSten and for ChSten. How might the data be modelled? A reasonable model turns out to be

$$\text{ChSten} \sim \text{Clinic} + \text{Trt}/\text{poly}(\text{InitSten}, 2)$$

Something new has been added!  $\text{Trt}/\text{poly}(\text{InitSten}, 2)$  means "fit a 2nd degree polynomial within each level of Trt". In words, therefore, the model is "Express ChSten as the sum of a Clinic

effect and a second-degree polynomial [including intercept] within each level of Treatment". This is not quite a covariance analysis — that would be accomplished by a model such as

$$\text{ChSten} \sim \text{Clinic} + \text{Trt} + \text{poly}(\text{InitSten}, 2)$$

in which the same second-degree polynomial would be used for each level of Treatment and Clinic. Because of the apparent outliers, it would be worthwhile to consider a robust analysis. Also, because the patients do not provide the same numbers of observations (lesions) a weighted analysis is needed, taking this into account. This is easily done in S-Plus:

```
attach(qcasall)
model <- ChSten ~ Clinic + Trt/poly(InitSten,2)
qcasall.glm.wls <- glm(model,gaussian,qcasall,nLesion)
qcasall.glm.rwls <- glm(model,robust,qcasall,nLesion)
```

The first call to glm fits the model by weighted least squares, the weights being supplied by nLesion. The second call to glm fits the model by robust weighted least squares. These analyses can be carried out remarkably easily with S-Plus.

The results can be displayed easily by using the summary.glm command and directing the results to a dataset:

```
sink("g:glm.out")
summary.glm(qcasall.glm.wls)
summary.glm(qcasall.glm.rwls)
sink()
```

Here are the key results, after some prettying up using a text processor:

```
Call:      glm(formula =      ChSten ~ Clinic +
              Trt/poly(InitSten, 2), family = robust,data =
              qcasall, weights = nLesion)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-6.27	-1.775	-0.3197	1.798	8.19

Coefficients:	Value	S.E.	t value
(Intercept)	1.741	0.341	5.103
Clinic	0.128	0.328	0.389
Trt	0.348	0.287	1.210
TrtBpoly(InitSten, 2)1	-8.524	7.098	-1.201
TrtApoly(InitSten, 2)1	-26.106	7.651	-3.412
TrtBpoly(InitSten, 2)2	22.778	9.452	2.410
TrtApoly(InitSten, 2)2	26.924	11.094	2.427

(Dispersion Parameter for Robust Gaussian family \_ 134.1225)

Null Deviance: 43967.68 on 219 degrees of freedom  
Residual Deviance: 1548.032 on 213 degrees of freedom

Continued on the next page...

## Correlation of Coefficients:

	(Intrcpt)	Clinic Trt	ply1B	ply1A	ply2B	
Clinic	0.541					
Trt	0.018	-0.032				
TrtBpoly(InitSten,2)1	0.077	0.117	-0.020			
TrtApoly(InitSten,2)1	0.153	0.067	0.136	0.008		
TrtBpoly(InitSten,2)2	0.178	0.025	-0.196	0.124	0.002	
TrtApoly(InitSten,2)2	0.220	0.018	0.249	0.002	0.206	0.0004

These examples sketchily illustrate some very powerful capabilities of S-Plus as applied to the analysis of real data arising in biopharmaceutical applications. By no means do they exhaust the possibilities with S-Plus, nor do they comprise the only analyses that might be carried out on these data. Finally, even

though 1500+ functions are coded into S-Plus, the flexibility of the language allows custom procedures to be programmed, too. Indeed, a large library of procedures for performing very sophisticated analyses exists, and procedures are available from this library for the asking via e-mail.

## ASA Biopharmaceutical Section Executive Committee—1993

### Immediate Past Chair

**Camilla A. Brooks, Ph.D.**

President, CB Quantitative, Inc.  
PO Box 659  
Upper Marlboro, MD 20773-0659  
(301) 868-1097  
(301) 868-4474 Fax

### Chair

**Bruce E. Rodda, Ph.D.**

Vice President, Biostatistics & Data Management  
Bristol-Myers Squibb  
Pharmaceutical Research Inst.  
PO Box 4000  
Princeton, NJ 08543-4000  
(609) 921-5776  
(609) 921-5740 fax

### Chair - Elect

**Robert R. Starbuck, Ph.D.**

Assistant Vice-President  
Clinical Biostatistics and Data Management  
Wyeth-Ayerst Research  
145 King of Prussia Road  
Radnor, PA 19087  
(215) 341-2070  
(215) 341-2092 Fax

### Secretary - Treasurer

**Robert L. Davis, Ph.D.**

Director of Biostatistics  
Astra/Merck Suite 300  
300 Berwyn Park  
Berwyn PA 19312  
(215) 651-7870  
(215) 651-7996 Fax

### Section Financing

**John Schultz, Ph.D.**

The Upjohn Company  
200 Portage Road  
Unit 9165-32-2  
Kalamazoo, MI 49001  
(616) 385-7427  
(616) 329-5548 fax

### Program Chair

**Mark Scott, Ph.D.**

Director, Biometrics & Medical Systems  
Clinical and Medical Affairs  
ICI Pharmaceuticals  
Office Wing 3  
Routes 202 & 141  
Wilmington, DE 19897  
(302) 886-8495  
(302) 886-2442 fax

### Program Chair Elect

**Kenneth J. Koury, Ph.D.**

Building 60, Room 203  
Lederle Laboratories  
North Middletown Road  
Pearl River, NY 10965  
(914) 732-3827  
(914) 735-5249 Fax

### Work Group Coordinator

**Nick Teoh, Ph.D.**

Biostatistics  
Shering-Plough Research Institute  
2000 Galloping Hill Road  
Kenilworth, NJ 07033  
(908) 298-5986  
(908) 298-4688 Fax

**Chair of the Midwest Biopharmaceutics  
Statistics Workshop**

Patrick D. O'Meara, Ph.D.

Director, Statistical Services Division  
624 Peach Street  
P.O. Box 80837  
Lincoln, Nebraska 68501  
(402)-476-2811  
(402)-476-7598 fax

**Representative, Applied Statistics Conference**

Karl E. Peace, Ph.D., FASA

President, Biopharmaceutical  
Research Consultants, Inc.  
4600 Stein Road, Suite B  
Ann Arbor, MI 48105  
(313) 663-4440  
(313) 663-7797 Fax

**Publications Officer**

Chris Gennings, Ph.D.

Biostatistics Department  
Medical College of Virginia  
Box 32  
MCV Station  
Richmond, VA 23298-0032  
(804) 786-9824  
(804) 371-8482 fax

**Biopharmaceutical Report Editor**

Avital Cnaan, Ph.D.

Division of Biostatistics, Department of Pediatrics  
University of Pennsylvania  
The Children's Hospital of Philadelphia  
34th Street and Civic Center Boulevard  
Philadelphia, PA 19104-4399  
(215) 590-4486  
(215) 590-4487 Fax

**Chair of Committee on Advisors to FDA**

Vern Chinchilli, Ph.D.

Associate Professor  
Department of Biostatistics  
Medical College of Virginia  
1 East Marshall Street  
PO Box 32, MCV Station  
Richmond, VA 23298-0032  
(804) 786-9824  
(804) 371-8482 Fax

**Council of Sections Representative**

Edward S. Nevius, Ph.D.

Division of Biometrics  
Food and Drug Administration HFN-713  
5600 Fishers Lane  
Rockville, MD 20857  
(301) 443-4594

**Council of Sections Representative**

Janet M. Beegun, Ph.D.

Vice President  
Pharmaceutical Product Development  
1400 Perimeter Park Drive, Suite 100  
Morrisville, NC 27560

(919) 380-2000  
(919) 380-2022 Fax

**Liaison Activities**

Helen Bhattacharya, Ph.D.

Director of Biostatistics  
Sanofi Pharmaceuticals, Inc.  
40 E. 52nd Street, 13th Floor  
New York, NY 10022

**Executive Committee**

John Lambert

Director, Biomedical Operations  
Sandoz Pharmaceuticals, Inc.  
Route 10  
East Hanover, NJ 07936  
(201) 503-6914

**Executive Committee**

Gary Neidert, Ph.D.

Director of Clinical Data Management  
9165-298-139  
The Upjohn Company  
301 Henrietta Street  
Kalamazoo, MI 49001  
(616) 329-8591  
(616) 329-5579 fax

**Executive Committee**

Akbar Zaidi

CDC Mathematical Statistician  
Division of SPD/HIV  
Center for Prevention Services  
1600 Clifton Road, N.E.  
Atlanta, GA 30333  
(404) 639-2562  
(404) 639-2555 fax

**Executive Committee**

Lillian Kingsbury, Ph.D.

Director of Biostatistics  
Bio-Pharm Clinical Services  
512 Township Line Road  
Blue Bell, PA 19422  
(215) 283-0770  
(215) 283-0733 fax

**Executive Committee**

Jerome Wilson, Ph.D.

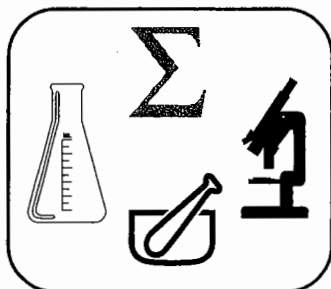
Director of Biostatistics and  
Data Management  
Warner-Lambert  
170 Tabor Rd., Room 3015  
Morris Plains, NJ 07950  
(201) 540-2422  
(201) 540-4300 fax

**Executive Committee**

Nguyen V. Dat, Ph.D.

Associate Director of Biometrics  
Smith Kline Beecham  
P.O. Box 1510 M/C FF0605  
King of Prussia, PA 19406  
(215) 270-6277

# Biopharmaceutical Section-Sponsored Sessions at ENAR Meeting, March 21-24, 1993, in Philadelphia, Pennsylvania



Don't forget to come to the Biopharmaceutical Section-sponsored Sessions at the 1993 ENAR Spring Meeting.

1. Adjusting treatment effects for baseline and other predictor variables.  
*Organizer/Chair: ROGER*

*E. FLORA, Pharmaceutical Research Associates, Inc.*

2. Design and analysis issues in clinical trials in epilepsy.  
*Organizer/Chair: LILLIAM KINGSBURY, Bio-Pharm Clinical Services, Inc.*

3. Measurement of efficacy with repeated measures and missing data. *Organizer/Chair: AVITAL CNAAN, University of Pennsylvania*

4. Surrogate Markers. *Organizer/Chair: BRUCE RODDA, Bristol-Myers Squibb Pharmaceutical Research Institute*

5. Dental Data Analysis. *Organizer/Chair: PETER B. IMREY, University of Illinois*

6. Post-marketing surveillance in the pharmaceutical industry: the roles of sample survey methodology and epidemiology. *Organizer/Chair: CAMILLA BROOKS, CB Quantitatives*

*See the complete program in the February 1993 Amstat News for more details.*

## Let's Hear from You!

*If you have any comments or contributions, contact:  
Dr. Avital Chaan*

*Division of Biostatistics*

*The Children's Hospital of Philadelphia*

*34th Street and Civic Center Boulevard*

*Philadelphia, Pa 19104-4399*

*(215) 590-4486*

*(215) 590-4487 Fax*

*The Biopharmaceutical Report is a publication of  
the Biopharmaceutical Section of the American  
Statistical Association.*

*© 1993 The American Statistical Association*

## Biopharmaceutical Report

c/o American Statistical Association

1429 Duke Street

Alexandria, VA 22314-3402

USA

FIRST CLASS POSTAGE

FIRST-CLASS MAIL  
U.S. POSTAGE  
PAID  
WASHINGTON, DC  
PERMIT NO. 9959

2419 12/93  
DR. ROBERT L. DAVIS  
215 MANOR ROAD  
HARLEYSVILLE PA 19438