Biopharmaceutical Report

American Statistical Association

Volume 1, No. 5
Fall 1992

Chair: Camilla Brooks, Ph.D.
Editor: Avital Cnaan, Ph.D.


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 total budget and effort to collect data entirely from a safety perspective. Safety data are generated from clinical adverse events, safety laboratory exams, and physiological tests such as ECG and 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 ensuring a patient’s safety in a trial, safety data allow 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 trend 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 (Bray, 1991), others may shun them. (Kramer, 1991). Many statisticians take the middle ground and use the p-values obtained from inferential statistical methods for exploratory purposes (Ast, 1987, 1990). In terms of analyzing lab data, Sogolow-Gilbert, Mobair and Subeloff (1986) proposed a method to combine related lab results to study the extent of functional variances reflected by the lab results. Similar to Sogolow-Gilbert, Mobair and Subeloff’s idea of using several parameters at a time, a vector-based re-sampling method was employed by Ramsey and Elias (1991) to study a cluster of adverse events within a body system. Realizing that all of the above efforts concentrate on one-blind-at-a-time analyses and safety data are really multi-faced, Chuang-Stein, Moberg and Mandelbaum (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 any meaningful, or is what we are doing ML?
Biopharmaceutical Report, Fall 1992

know what we are doing enough? Are we doing the right things? 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 & facilitate the treatment selection when safety is the primary consideration? I have often asked myself these questions. The answers, a physician should be taught 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 differences outside 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, stimulate further discussion on this topic. I hope this article provides 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. Eventhough 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/therapy and to study its overall safety profile. Because of the limited size, a trial that enrolls only dose and effect data without the knowledge of high incidence and the typical frequency listing of adverse effects with their intensity 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 in linking together findings from separate analyses to produce an overall safety profile. To produce the overall safety profile, one needs to resort to all the safety information and combine each information from all sources. This need exists for all pre-marketing trials, and especially so for phase II clinical trials when the experience with the drug is limited and one is on a postmarketing journey.

Since safety information comes from multiple sources that can be equally important, the simultaneous analysis of such information resembles that of the efficacy data with there exist multiple efficacy endpoints. While multiple efficacy endpoints have received a sensible attention (Reynolds, 1992; O'Brien, 1994; Pocock, Geller, and Tzartos, 1994), Geller, Galsom, and Goldschmidt, 1989, etc.), multiple safety variables have not. Chiang-Stein, Mombert, and Masserman (1992) proposed to structure the massive safety data into a more manageable framework by ascertaining 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 such classes, they propose to assign to each patient an overall intensity grade based on all relevant information. The analysis of such organized data concomitantly on a simultaneous comparison of the mean intensity grades for different treatments within each class with the use of a multivariate analytic and scores that reflect the acceptability of the various intensity levels. One advantage of their approach is its ability to refine 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 grade-off is the decision rule that one needs to determine beforehand to culminate the information. On the other hand, the decision rule forces one to think hard in advance about what constitutes a more accurate safety concern for a target population. Such an exercise also mimics for practitioners 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 interventions) between the two treatment groups. At times, analysis by baseline of events is conducted. Such analysis examines the occurrence 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 close disagreement 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 those parameters. Other such occurrences, such as abnormality observed on ECG or change in 24-hr, are typically compared using binary outcome techniques (with or without adjustments for covariates).

All the above analyses are informative in their respective roles. There are various approaches to handle the multitude of references for 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 glanced. Worse yet, many the comparisons appear to be data-driven. The latter partially motivates from Fidels (1987) requirement that rigorous statistical methods be applied to every 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 III trials, safety analyses for phase IV trials carry a different mission. Since phase IV trials are the informal testing where treatments are compared directly, they provide the essential information required for the form based on both efficacy and safety consideration. As a result, safety analyses for such trials go beyond the identification of s,t; more summarization of adverse experiences.

Interventions are given with the hope to cure or control, presenting different types of challenges. Moreover, in the case of serious adverse events, there is also the danger that an intervention can cause harm or injury. Depending on the distress 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 life-threatening. In those cases, evidence for treatment-induced harm is low. For the other hand, safety data for treatments-induced harm is high for idiosyncratic diseases such as cancer and AIDS. This intricate relationship between benefit and risk led to the general belief that there 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 recognizability of the comparable roles of benefit and risk in the evaluation of treatments, the analysis of clinical trials will lead to a great variety of benefit-risk assessments with risk estimated by the safety data. The pharmaceutical industry quantitatively separates efficacy and safety summaries in a new drug application, and it is inconceivable that a serious effort is launched in the same application to include both endpoints in a joint analysis. As a result, one treatment provides more efficacy
Biopharmaceutical Report, Fall 1992

with a pint tag of more side effects, the decision on the treatment
advises becomes a difficult one. Some formal statistical tools are
definitely needed to address the simultaneous benefit/risk
currently are weighed by the drug developer and
Sinkhole (1991) made an attempt in this direction using ordinal
response data. Statisticians need to give more thought to the
practice of monitoring 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'Niel (1998). 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
systemic and genitourinary side effects following the use of X-ray
exposure in contrast medium. Kessler and Friedman (1975) noted
that, in addition, rare side effects start to show up during the
post-marketing epidemiologic follow-up studies. I highly recommend
continuing 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
treatment religiously pools data from different studies to create the
required safety summary. A typical approach at the moment is to
collate data from all studies and summarize them as if they
came from a single study. However, is this approach appropriate?
Differences in treatment plans and target populations can introduce
extra variance for parameters of interest, but the current way to
quantify the diagnostic safety summary does not include a provision for such
a variance component. Even in the absence of the variance term, one can
mostly imagine 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?
Considering 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-analytic 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
parameter in the sample is used because of the wide range of
applications. This approach incorporates a variance component
that addresses the fact that the data in each study exhibit a between-study variability.
The excess 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 requests 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 stems 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 specimen are being processed at the
respective study site 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 prescreen normal population
misses out fundamental issue, i.e., whether the new drug/treatment has 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 treatment 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 as a clinical trial. First, the group of patients
participating in a clinical trial use assays intended for diagnostic
evaluation of a disease state, and not for assessing 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 need to scrutinize is generally true 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 term. 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 that better fits the lab assay are traditionally
used to diagnose a disease state, a laboratory tends to balance the metrics in such a way that the existing biochemical 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 needed at the extreme of the normal range is not overly
stringent 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 the 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 (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
prescreened in a controlled manner, van Straalen and Chauang-Stein (1993)
proposed to use a study's inclusion/exclusion criteria to define
a reference population. In addition, they recommended using
patients' pre-treatment lab results adjusted by a laboratory's
performance in a proficiency survey to construct a set of sol-ly-
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 abnormal lab performance by having all enrolled laboratories analyze a common set of specimens. After analyzing the specimens, laboratories and data go to one coordinating center which collates and compares 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 interpretaton. Oliver and Chuang-Stett (1993) recommended a two-clone adjustment procedure that adjusts lab values based on a laboratuty'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 patients: 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 same process of data from different laboratories within one study where differences occur as a result of laboratories using different procedures or equipment.

When analyzing lab data, it is a common practice to generate a table that flag lab values (at all evaluations points) that are outside the reference ranges with the ranges being more supplied by individual participating laboratories. Since the current reference ranges are constructed based on data from a normal population, the practice does not make much sense for situations where patients enter a trial with severe functional impairments. In addition, such practice infers nothing as to how a patient's biochemistry was affected by the treatment. But then, why are we still using it and why it is still being routinely requested 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 derived based on data from all sources. This requires effort to intelligently combine all safety information to facilitate the overall summarization. The effort takes work, but the work makes one use the puzzle together to create a more succinct picture of the safety outlook of a new intervention.

One thing that I have not touched upon but agreeable requires an equal amount of thought, is the collection of the safety data itself. Because they usually do not exist any definite forms when being collected as medical events, they 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-ended or using a checklist? Should they be volunteered? How should multiple episodes of the same event be handled? Should there be a change in the intensity of a clinical event be recorded? Fairly little guidance has been given over time to some of these issues if we are trying to spend so much effort and time in a trial to collect safety data. Should we do it as consistently as we possibly can and make the best use the data allows us?

References


Biopharmaceutical Report, Fall 1992


Discussion

Quality Safety Information that Meets Customers Needs.

Gregory G. Enas

Lilly Research Laboratories, Eli Lilly and Company

Dr. Chryse Chuang-Stett 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 case in point of many of the current issues and has helped to coordinate and solidify my own thinking about these differences. 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 over have been an afterthought on the analysis side.
As Christy has ably pointed out, safety is truly a multivariate characterization which allows it to be the most ill-defined entity in the benefits/risk ratio. Yet, previously no work on cost-benefit methodology and safety adjudication was a practical attempt to analyze 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 medications 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 instances that have been undertaken to address the needs of regulators. The insured 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 investigational sites. Use of central laboratories to handle all aspects 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 standardized event outcome measures, such as "Vomiting" Emergent Signs and Symptoms (TESS), also help expedite the speed of the development of a new medicinal regimen (Ossen, 1985). Many of these innovations we have seen come to pass for premortem safety analyses may play an important role in the expansion of postmarketing surveillance studies and prospective complaints surveillance.

Another key customer are the investigators who conduct the studies. In particular, much of the premortem 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 bench on animal toxicology. These rules take the investigator's safety concerns and uncertainties and translate these into a frequency which will allow one to distinguish a particular compound 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 in a priori. The design of these studies could then take these events into account when establishing safety and the standardization of safety data collection.

Christy's suggestions on the integrated safety summary as an exercise which would be valuable. To take this one step further for the practicing physicians, however, we have not seen evidence to date of package inserts being crafted from carefully conducted meta-analyses. For fun I reviewed all of the drug associations 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 type studied, investigator differences in 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 constructing an integrated safety summary, this would also help. For events with relatively high frequency, meta-analytic thinking could help sort out different 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 similarly as possible, without obfuscating any individual adverse events, drug interactions, or patient prognostic characteristics (e.g., elderly versus non-elderly patients, photosensitivity). Gehan (1991) and Lilin (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 pharmacist review a study by collecting and analyzing safety information on an ongoing basis throughout a study. In a similar manner, statisticians can contribute to improve their working relationships with study physicians and monitors, FDA scientists, and other regulatory personnel in order to standardize collection, analyses, 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 included in the label which allow one to distinguish a particular compound 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 in a priori. The design of these studies could then take these events into account when establishing safety and the standardization of safety data collection.

Biopharmaceutical Report, Fall 1992
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. Chua
Chung-chia, that 70-80% of the effort expended to data
collection associated with clinical trials of new drugs concerns
storage. Although this may seem rather negative, the opt-out
option is often the most important issue to consider
when assessing whether to pool or not. This is because
for several reasons, the data are still very strong.

Several years ago the most important issue to consider
was the importance of post-marketing surveillance studies. Some
interests in this regard are given by Tsong (1992), Kadowaki and
Sporn (1988), Rawson et al. (1990), and Sarbajana and Colburn (1988).
New emphasis on
using innovative quality-control measures is of interest to
post-marketing safety data.

Summary
Dr. Chung-chia’s paper about the analysis of safety data tends
to also consider the important issues of who needs this
information and what can be done prospectively to get the right
information to them. This means that statisticians must play
a stronger role 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
the use of quality information, rather than using collection in
these data studies, using use of specific reference ranges for
laboratory data, pooling different studies into integrated safety
summary reports, and so on.

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

References
techniques for the surveillance of birth defects through a Monte
Carlo simulation, Statistica Medica, 3, 239-247.

haematological toxins in a Meta-Analysing, Biometrics, 47, 899-906.


for postmarketing surveillance of adverse drug reaction reports,
Statistica Medica, 7, 103-1030.

Proceedings of the Biopharmaceutical Section, 14-16.

O’Neill RT (1988). Assessment of safety, Chapter 13 in
Biopharmaceutical Statistics for Drug Development (Ed. Karl E.
Preis), New York: Marcel Dekker, Inc.

monitoring: methodology and recent progress, Journal of
Clinical Epidemiology, 43, 503-522.

Tsong Y (1992). False alarm rates of statistical methods used
to determine increased frequency of reports on adverse drug

early stopping rules for a clinical trial based on observed
numbers of adverse events.
Biopharmaceutical Report, Fall 1992

Table 1: All Adverse Experiences - CATS

<table>
<thead>
<tr>
<th>Treatment Group</th>
<th>AEs</th>
<th>Probably Not Related</th>
<th>Unknown</th>
<th>Probably Related</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLACEBO</td>
<td>127</td>
<td>156</td>
<td>93</td>
<td>3</td>
</tr>
<tr>
<td>Rate/100 Assess.</td>
<td>4.1</td>
<td>5.0</td>
<td>3.0</td>
<td>2.0</td>
</tr>
<tr>
<td>TICLOPIDINE</td>
<td>176</td>
<td>339</td>
<td>275</td>
<td>9.9</td>
</tr>
<tr>
<td>Rate/100 Assess.</td>
<td>6.4</td>
<td>12.1</td>
<td>9.6</td>
<td>7.5</td>
</tr>
<tr>
<td>Rate T/Rate P</td>
<td>1.6</td>
<td>2.4</td>
<td>3.3</td>
<td>1.5</td>
</tr>
</tbody>
</table>

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

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 conducted 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 sub-acute 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 FDA regulations. Patients were followed at essentially 4-month intervals for an average of 24 months. The AEs section of the CIBF was fairly typical and asked clinicians to judge the relationship to study drug as "probably related", "probably not related", or "uncertain". Ticlopidine proved to be efficacious, but like many active drugs had "active" side effects. In Table 3, we have summarized the rate of all AEs reported separately for placebo and Ticlopidine, subdivided by the likelihood of being drug related. The rate of the frequency of reports was over 5 to 1 for the subgroup judged as "probably related", and over 10 to 1 for those judged "probably not related", with the "unknown" group falling in the middle with a ratio of 3 to 4.1. 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.

In conclusion, the data show that Ticlopidine is efficacious in reducing the risk of subsequent vascular events in patients who have had a stroke. The rate of adverse events related to study drug was lower for Ticlopidine than for placebo, with the "unknown" group falling in the middle. The data support the hypothesis that clinicians can sort out real adverse reactions, although the ratio of 1.6 for the unrelated group indicates that some are overlooked.

Objective of Safety Analysis

a. Phase I and II

The method proposed by Dr. Chuang-Stein for competing the weighted intensities of adverse effects in a single multivariate comparison is ingenious and poses great strength. Being of the class of statistics that regards p-values as largely irrelevant in the context of safety, countervails Table 2 shows the same type of approach as statistical overkill. However, in certain situations, this 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 III stage of drug development, I believe the key objective in safety analysis is to characterize the adverse effects of a new drug. In this perspective, vice 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 biologically about the action of the drug, and others which are unexpected idiosyncratic. Rapid recognition of extremely troublesome side-effects may allow the early termination of the drug trial 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 III 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 textual decriptions of similar events into a common set (not necessarily always from within a single body system), a task 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 judgements.

Table 2: Rash in CATS

<table>
<thead>
<tr>
<th>Treatment Group</th>
<th>AEs</th>
<th>Probably Not Related</th>
<th>Unknown</th>
<th>Probably Related</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLACEBO</td>
<td>39</td>
<td>16</td>
<td>7</td>
<td>0.51</td>
</tr>
<tr>
<td>Number Reports</td>
<td>1.23</td>
<td>0.51</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>Base/100 Assess.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TICLOPIDINE</td>
<td>33</td>
<td>36</td>
<td>42</td>
<td>5.31</td>
</tr>
<tr>
<td>Number Reports</td>
<td>1.19</td>
<td>1.37</td>
<td>4.31</td>
<td></td>
</tr>
<tr>
<td>Base/100 Assess.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rate T/Rate P</td>
<td>1.0</td>
<td>2.7</td>
<td>6.8</td>
<td></td>
</tr>
<tr>
<td>Treatment Group</td>
<td>Definitively Related</td>
<td>Possibly Related</td>
<td>Not Related</td>
<td>Possibly Related</td>
</tr>
<tr>
<td>------------------</td>
<td>---------------------</td>
<td>------------------</td>
<td>-------------</td>
<td>------------------</td>
</tr>
<tr>
<td>CYCLOSPORINE</td>
<td>23</td>
<td>196</td>
<td>302</td>
<td>215</td>
</tr>
<tr>
<td>Rate/100 Cases</td>
<td>8</td>
<td>32.0</td>
<td>45.3</td>
<td>35.1</td>
</tr>
<tr>
<td>PLACERBO</td>
<td>1.2</td>
<td>17.5</td>
<td>211</td>
<td>31</td>
</tr>
<tr>
<td>Rate/100 Patients</td>
<td>67</td>
<td>17.5</td>
<td>211</td>
<td>31</td>
</tr>
<tr>
<td>Rate/100 Aspects</td>
<td>11.2</td>
<td>1.7</td>
<td>1.4</td>
<td>6.8</td>
</tr>
<tr>
<td>Rate/100 P</td>
<td>0.3</td>
<td>1.1</td>
<td>1.4</td>
<td>6.8</td>
</tr>
</tbody>
</table>

### b. Phase III

The output of the controlled fish farming experiment of Phase III safety data analysis is hopefully a relatively small set of characteristic drug reactions associated with an investigational drug that can range the safety review of Phase III. General reviews of adverse experiences are also required but the potential, if desired, for formal hazard 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 overall trend of therapeutic benefits and risk as essentially subjective. The basic problem is not having a common quantitative unit with which to bring together benefit and risk. Our occasional 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 payers. It is quite possible that these different criteria will reach different conclusions about the relative merits of competing agents, as it is happening currently with the various thrombolytics available. Clearly in some situations total morality is the only relevant endpoint and naturally combines at least part of the benefit/risk profile. Occasionally, a single major adverse experience may be directly combutable with the primary efficacy outcome to reflect the net position. Examples of situations might include 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, in part the way that the information is collected in clinical trials, although by definition unblinded and usually uncontrolled. In other words, individual patients have to be recruited to use as sentinel 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 jaded view of the validity of post marketing surveillance is colored by an earlier experience with the drug Synthalin. The plant lyceu agents had been marketed in Europe for a number of years when I was conducting a North American based trial [Jenn, 1985] in stroke patients. It did not take me long to become aware of Synthalin’s hypothermic and, in the face of a very modest trend towards efficacy, the linear analysis, decided to stop the trial early. The lack of appreciation of the problematic early trend toxicity prior to our study was one of our most striking lessons when reviewing the safety data findings. An evaluation of available post marketing data in Europe did reveal a clear indication of the problem. This re-evaluation lead quite rapidly to the worldwide withdrawal of the drug. While this anecdotal experience may have less relevance in North America, it does support, I think we all agree, that statistical analysis of safety data is much safer if we know what we are looking for.

### Integrated Safety Summaries

As an academically-biased statistician with experience in clinical trials, I have yet to be involved in formally integrating data over studies and thus only on the efficacy side of the equation. The various forms of meta analysis have gained great popularity following Peet’s [1980] attempts to extract mean out of the apparent dearth 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 variability simply because of the open-endedness of the resulting 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 one good no matter how good it is. In general, I prefer to summarize laboratory data quantitatively rather than qualitatively. One might argue that this gives one more prominence, for instance, in liver function tests. It is possible that these different criteria 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 endpoint and naturally combines at least part of the benefit/risk profile. Occasionally, a single major adverse experience may be directly combutable with the primary efficacy outcome to reflect the net position. Examples of situations might include ischemic stroke plus infarct related 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.

### Concluding Thoughts

The most direct in my mind is that the open-ended nature of safety data creates an inherent challenge in individual studies, let alone in the bringing together of many studies into an integrated synthesis. Dr. Chang-Stein’s paper is an honest appraisal of the difficulties which precludes to possible solutions while ten glancing 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 exclude the concluding remarks of making "all plans out of one's own“ but this is too pessimistic a note on what I hope to do. This is a note on what I hope to do.
**Discussion**

David Salsburg  
**Pfizer Central Research, Pfizer, Inc.**

Dr. Chung-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 re-read) the Chung-Stein, Mobberg, and Musselman paper in *Statistics in Medicine.* In this paper, the authors consider many of the problems in safety analysis with a simple device. They note that the medical professionals examine the data and categorize each patient in terms of severity of adverse events by a "system." As a result, the analysis deals with a well-defined, specifically important set of events.

Very careful analysis of 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 spans patients are followed for periods up to three years, across dose (which may change over time), across types of observations (break side effects, signs of disease, blood chemistry, etc.), across body systems or syndromes, and across patient types (male/female, aged/pregnant, black/white, etc.). It is impossible to display events across all five of these dimensions, and so we slice 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. Chung-Stein points out, it makes no sense to equate statistically significant changes in blood chemistry with lack of safety or the lack of statistical significance with safety. Most drugs are nephrotoxic 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 stimulatory, such as increase in one liver enzyme and decreases in another. Dr. Chung-Stein points out that the "normal ranges" are notions that do not apply to normal distributions across patterns and are often based on last observations in addition to standard units 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 cells and the lymphocytes drop well below their normal range. Drugs that involve liver toxicity cause SGOT to rise 10 to 100-fold.

Contrast this with the NSAIADrugs, 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. These may be long-term consequences of these changes, but these cannot be evaluated in the context of a new drug development program. It is a good idea to think of the potential danger of a new drug in terms of 24-42 table where the rows represent:

- Events that might be expected from the drug's pharmacology may be observed in patients with rheumatoid arthritis. Levett 335:1051-1055.

- Events that are not observed.

- Columns represent:

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

Examples of high frequency events that might be expected from the drug's pharmacology are podal edema with a calcium channel blocker, or pseudohypertension 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 demographic doses. An example of a new event that might be expected is the interaction between MAC 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 moribund 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 bogie 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 due to local laws, or they will occur so rarely as to be of no consequence whether they are drug-related. Unless society wants to stifle all future development, drug safety, society must learn to live with that uncertainty.

Thus, safety analysis consists of what can be done with the data and the drugs. The expected high frequency changes must be characterized through the use of logistic regression 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 physicians away from the drug. The statistical program addressed by much of Dr. Chung-Stein's paper deals with techniques for locating the relatively high frequency unexpected event. I do not have any better answers than the does.

**Discussion**

R. Srinivasan  
**Division of Biometrics**  
**The Food and Drug Administration**

The paper by Chery Chung-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 derived 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 compounds because that's what we need to know in the label for physician's use.

Another important issue in the analysis of adverse events is multivariate. 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<0.001). All and others will have higher p-values. Braas (1980) and Hans (1991) have noted that a higher p-value (e.g., p>0.05) does not necessarily imply lack of association between treatment and an adverse event because of relatively low power that is available to detect such associations with the available sample size in more clinical trials (or their meta-analysis combinations). According to Good (1991), for any study, an important role of analysis 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 Sinkula (1990), has proposed three ways to incorporate benefits and risks into one analysis. The first extends Hildreth (1990) procedure while the others generalize the benefit/risk ratio considered in Paine and Lukan (1990). All these procedures employ weights to compute the summary statistics. The author's concept of benefit-risk assessment is to be made with respect to the underlying disease or symptom is also reflected in Japanese Protocols as a 'usefulness' criterion to be assessed by the investigator. The investigator recommends conducting additional benefit-risk assessments in the presence of newly detected side effects after a new drug/therapy has entered the market. In general, this is not practical to accomplish in the regulatory rules.

The author advocates that as in efficacy evaluation, meta-analysis can be carried out to pool safety data from different studies. He points out that when pooling data from different studies, one has to pay attention to the nature of the problem, data retrieved 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. Survival evaluation procedures such as logistic regression with covariates may be used to adjust inferences. In the same way, the author advocates for the use of the disease estimate in drug development.

Laboratory results provide very useful information regarding a patient's safety in a trial. Lab-specific ranges are commonly used in multi-center trials where patients' 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 one of "normal range", and (2) disease averse laboratory experiences a adverse clinical experience interaction. The author strongly feels that safety monitoring should focus on assays in the opinion of the co-pre-existing range in a patient's biochemistry. It is a fact that the author is not aware of a universal definition of reference populations and a common standard to establish reference ranges. The author feels that these two deficiencies along with the differences in assay procedures and the accuracy in diagnosis raise doubts about the usefulness of reference ranges in multi-center 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 comparing reference ranges based on data from normal populations is not appropriate when laboratory results of bexty systems are affected by drug or disease. I would think that one measure of therapy benefit is far from valid and so 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 Herbes, my supervisor and my colleague, and Ms. Beth Turner for offering useful suggestions in preparing this discussion.

References

Discussion
Janet Wittes
Statistics Collaborative
Washington, DC 20036

My introduction to the analysis of safety data in clinical trials was during my first days as Safety Monitoring Board Meeting. Confronted with long lists of laboratory parameters, symptoms, signs, and events, I did what any rational person would do—I asked Mary Halpin. 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 sighed and gave me sage advice, "Tone out during the discussion." After ten years of ruminating, I welcome Dr. Chuang's thoughtful discussion on how to collect, analyze, and interpret safety data. He so ably points out that safety should 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 with other ways? As he says, "an overall safety profile can be drawn based on data from all sources." This requires an effort to intelligently combine all safety data to facilitate the overall summarization.

I find stimulating previous work with Mohberg and Sinkula (1990) integrated safety and efficacy variables. Her current paper discusses a number of important issues, 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 range" in the context of clinical trials on people with disease. Limiting all the discussion is the problem of assigning medically meaningful levels of importance to the many
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. Shenner 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 he sees is that "clinicians must regain control over clinical trials." Dr. Shenner properly condemns the inappropriate use of such things as statistical methods, analyses of convenience, and inadequate consideration of objectives. However, there are not failings that can be broadly attributed to statisticians but must be shared by all principals in trial.

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

Dr. Shenner states in this example that "we are not encouraged to carefully model..." and "We are (in)dicated to focus on a drift and featureless null model." We must question who is "not encouraging" and who is "encouraging" and, moreover, why, since this is true. A good statistician does not act in such a manner,

The "test the null hypothesis" mindset set is, unfortunately, a pervasive disease that infects both clinicians and statisticians. But it is improper to identify the mind set with the disease, because hypothesis testing has 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. Intent 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 conduct a test, depending only on univariate analysis. Dr. Shenner also expresses a concern regarding statistics in restricting a clinician's ability to modify a protocol by looking at a 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 a manner that allows clinicians to create large and immeasurable bias. No statistician 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. Shenner, statisticians have done little to educate clinicians about basic statistical concepts. We must certainly accept responsibility for this shortcoming, and we are all and willing to do whatever is necessary to correct this situation. But to be successful, we need the commitment and support of medical schools in two areas. First, bioethics and clinical trial methodology 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 educated 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. Shenner's article, "Reassessing Epidemiologic Authority," is right on target with one exception, and that is the statement with which we begin this letter: Clinicians must regain control over clinical trials. Clinical trials require a multidisciplinary approach, it is inappropriate for a clinician to direct a clinical study without appropriate management. 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 accessible as designed. It is not just want to do the 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 being to experimental therapy without a clear understanding of the likelihood of success.

Clinical 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 together 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
Canilla Brooks, PhD, Chair
Gladye Reynolds, PhD, Past Chair
Biopharmaceutical Section
American Statistical Association

References

measures made in assessing safety. What is reasonable calculus for assigning weight 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 assessments of the net complex of events as individual experiences (Pollman, Wais, and Cuker, 1992). For adverse events and safety data, the combination of substantial findings and clinical events is so complicated that I am pessimistic about finding generally applicable approaches that ameliorate unnecessarily charactirty experience.

Before I address two specific issues, Chang-Stein adds to the first topics, the problem of how to deal with the fact that adverse events and safety data contains many unmeasured 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 makes true adverse effects. Rare adverse effects 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 confuses the apparent adverse experience or drug. An elderly person who falls and breaks her hip in a clinical trial may show up in tables under "falls," "fracturing episodes," "dizziness," "hypertension," "euphoria," and "light-headedness." 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 feel well-aware that a single event can generate curiosity 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. Chang-Stein's various suggestions concerning multivariate approaches are useful: my own approach to categorizing variables is not by organ system, but by "syndromes" as well. Thus, each cluster of events as well as each 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 Chang-Stein's characteristics methods for sensitivity analyzing data on safety after a study is over as well as approaches for combining results from one or more clinical trials with results from nonclinical studies. At that point, there is often available a relatively static set of data that includes reliable information on efficacy. In 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 safety. Data and Safety Monitoring Committees usually have basic reproducibility. The simplicity of the tasks 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 the course of the study if the new therapy is sufficiently harmful to recommend continuation of the study or alternation of the dose of study agent. Many of the problems that Chang-Stein discusses are even more acute in the process of data monitoring because the data on adverse events and safety are so sparse and become the information on efficacy, which would allow at least an informed balancing of benefit against risk, are almost unavailable. During monitoring, the question of do-values is often moot. If the adverse event is harmful enough, we may not want to continue a study long enough to observe harm. On the other hand, if the safety data show minor laboratory abnormalities in a serious disease, we may be well advised to then statistically significant result is not sufficiently worrisome to necessitate changes. And sometimes, such as in trials of immunosuppression, early adverse events herald efficacy.

As Chang-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 for this complexity, physicians and statisticians should jointly design collection methods for safety data, and approaches to summarizing those data, and interpret the results.

References

Letter from the Editor
This issue of the Biopharmaceutical Review began with...
Review 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 aimed at 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 statistic series, this book is better suited to statisticians than to other scientists. The mathematics is extensive and complex. 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 strength 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. Rather than, most of the book is devoted to evaluating whether that shift in location is small enough that the two formulations can be judged as bioequivalent. There are two chapters on design, including a chapter on alternatives to the standard crossover design. Here 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 judicious this is 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 finding the evaluation of bioequivalence into the hypothesis testing framework. This leads to some awkwardness or incorrect questions 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 confidence interval. Although testing a null point hypothesis has little application in bioequivalence, the authors only discuss AUC 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 must other authors call "sequence effects." The book should at least relate the two. On page 19 and other places the authors repeat the common mistake of justifying the log-transform by the skewness of the observed data (e.g., AUC); it is the distribution of the model residuals that is important. On page 153 the authors report a common criticism of the Anderson-Hauck procedure, neglecting to point out that since bioequivalence data are not normally distributed 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.

Reference

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. These of the sessions are summarized below.

Longitudinal Analysis and Repeated Measures

George W. Divine

Penn Cereal Health Systems

Most of the papers in this section offered solutions to problems presented by missing data in longitudinal studies. The first presenter was Mrs. Jean J. Baker from the Laboratory, 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 's well presented, as the most of these three papers. The rest of the UNC presentation was given by Bernard Heister, who was also a co-author of the other two UNC papers. His talk entitled "Intention-Intolerant Incompole Repeated Measures Designs in Clinical Trials," described how power his 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 Repeated Longitudinal Data." He presented graphical representations of the fit of various convenient structural covariance models. Paul Von Tress of Acad 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 statistical issues. The fifth presentation was given by Ann Marie Boley 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 missing data distributions that will result in the indicated circumstance, and simulations were presented that suggested a fractional degree of freedom X2 distribution may give a good approximation for the likelihood ratio test statistic. Finally, Sandra Staggsee, 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 the situation. In short, the session gave some interesting results concerning some difficult questions in longitudinal analyses.
Biopharmaceutical Report, Fall 1992

Issues in Dose Response and Drug Combination Studies

David M. Lansky, Ph.D.
Seattle

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 centered groups on the type I error rate. For those who use the CA test regularly, the written report should be quite useful. The final talk compared two versions of a modified CA test (the modifications allow for something of a covariate). Pairwise prevalence method, and logistic regression on several data sets. While the p-values from the two methods are generally similar, the modified CA test method appears 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 quantification. The calibration intervals are markedly affected by the choice of the variance function. While it 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 concentration model for development of two mixtures. The model shows transitions from each of four states (tumor-free, tumor A only, tumor B only, and both) to death and transition to rates 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 get a scale-massantti property. The major point of the presentation was that the application of Lemma'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.
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 possess largely unique recognizability 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 narrate 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 attract attention. 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., 1750 World Trade Avenue North, Suite 500, Seattle, WA 98109, (206) 283-9822.

Example 1: Relating visual field threshold sensitivity (expressed as the logarithm of the threshold sensitivity measurement averaged over the 10 values observed 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 (T5 threshold sensitivity), L5 (lens density), and Age. The data started out as a simple ASCII file like this:

```
26 0.52 2.150
21 -0.18 2.475
71 0.81 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 would be easy to add the names to your data using commands like:

```
Age  L5  LogMean
26  0.52  2.150
21 -0.18  2.475
71  0.81  1.762
72  0.67  1.650
59  0.62  1.875
```

Support these data are in the data file "g-wisflida.dat." They can be brought into S-Plus as a table frame (in at least 6 rows, 3 columns) by:

```
> vflids <- read.table("g-wisflida.dat", 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 vflids, execute the command:

```
attach(vflids)
```

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

```
vflids.by <- kernel(smooth(Age, LogMean, bandwidth=15))
```

creates the smoothed values; add them to the plot with:

```
lines(vflids.by) to go.
```

A quadratic or cubic in Age might fit LogMean well. Similar steps suggest that LogMean is linearly related to L5. At worst, therefore, LogMean might be related to Age and L5 by:

```
LogMean \sim Age + Age^2 + Age^3 + L5 + L5^2 + Age \cdot L5
```

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

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

```
vflidsStage <- (Age - mean(Age))/sdev(Age)
vflidsStageL5 <- (L5 - mean(L5))/sdev(L5)
```

```
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 visitals:

```
visitals$tage2 <- tage^2;
visitals$tage3 <- tage^3;
visitals$tld1 <- tld/2;
visitals$tagedlt <- tage*tld
```

A list of the variable names would be convenient here:

```
vnames <- c("tage","tage2","tage3","tld","tld2","tagedlt")
```

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

```
xvars <- visitals[vnames]
```

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

```
allregs <- leaps(xvars, visitals$LogMean, labels=vnames)
```

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

```
attach(allregs)
```

```
attach(allregs)
```

Construct explicit output matrix

```
allregs.out <- cbind秭ure$labels)
```

```
sink()
```

Direct output to DOS file

```
sink()
```

Here are the key results from the exploration of all possible regressions: (Good models have Cp < 5)

<table>
<thead>
<tr>
<th>No.</th>
<th>Fit</th>
<th>prodxunm</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.13</td>
<td>old</td>
</tr>
<tr>
<td>2</td>
<td>45.72</td>
<td>tage</td>
</tr>
<tr>
<td>3</td>
<td>7.09</td>
<td>tage; tld</td>
</tr>
<tr>
<td>3</td>
<td>10.09</td>
<td>tage; tld</td>
</tr>
<tr>
<td>3</td>
<td>11.39</td>
<td>tage; tld</td>
</tr>
<tr>
<td>4</td>
<td>2.57</td>
<td>tage; tage2; tld</td>
</tr>
<tr>
<td>4</td>
<td>5.29</td>
<td>tage2, tage3, tld</td>
</tr>
<tr>
<td>4</td>
<td>7.87</td>
<td>tage; tage2; tagedlt</td>
</tr>
<tr>
<td>4</td>
<td>4.08</td>
<td>tage; tage2; tage3; tld</td>
</tr>
<tr>
<td>4</td>
<td>4.45</td>
<td>tage; tage2; tage3; tld</td>
</tr>
<tr>
<td>4</td>
<td>4.46</td>
<td>tage; tage2; tld; tld2</td>
</tr>
<tr>
<td>5</td>
<td>5.43</td>
<td>tage; tage2; tld; tld2; tagedlt</td>
</tr>
<tr>
<td>5</td>
<td>5.80</td>
<td>tage; tage2; tage3; tld; tagedlt</td>
</tr>
<tr>
<td>5</td>
<td>6.26</td>
<td>tage2; tage3; tld; tld2; tagedlt</td>
</tr>
<tr>
<td>6</td>
<td>6.29</td>
<td>tage; tage2; tage3; tld; tld2</td>
</tr>
</tbody>
</table>

These simple, easily executed analysis suggests a sensible, intuitively attractive model:

```
par(mfrow = c(1,3))
plot(visitals$LogMean ~ tage + tage2 + tage3 + tld, visitals)
```

and construct diagnostic plots

```
par(mfrow = c(3,1))
plot(visitals$LogMean)
qqnorm(residuals(visitals$LogMean));
qqline(residuals(visitals$LogMean));
```

Ordinate of first plot is LogMean, ordinate 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 anittenoma study

Kind of data supplied (with variable names added)

```
Alloc Clinic Tin nLesion Inistim twodim Orstem
1 B A 11 35.93 42.27 7.92
3 B A 2 40.56 40.90 0.34
5 B A 7 34.08 35.77 1.69
```

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

The relationship between OXstem and Tin, Inistim, and Clinic is what we are after.

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

```
attach(qqual) # qual = table frame with data
boxplot(oxstem~Clinic*Tin) # oxstem ~ T, match=T, outline = F)
yields
```
The distribution seems to have heavy tails: the lines go out 1.5 times the interquartile range. Also, the observations have differing precisions, as indicated by the values of n(Sten). 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 ln(Sten)? Plot it with the reciprocal.

qcasall.lm ~ lsmeans(ChSten, ChSten, bandwidth = 30)
plot(lm(ln(Sten, ChSten, xaxx = "e", yaxx = "e")

lines(qcasall.lm).

Looks like a quadratic at least — also, there are outlying values for ln(Sten) and for ChSten. How might the data be modeled? A reasonable model turns out to be

ChSten ~ ChSten + Trt + poly(ChSten, 2)

The formula for the model is: ChSten ~ Trt + poly(ChSten, 2)

A second-degree polynomial (including interaction) within each level of Trt. This is not quite a "converse analysis" — that would be accomplished by a model such as

ChSten ~ Trt + poly(ChSten, 2)

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

attach(qcasall)
model ~ ChSten + Trt + poly(ChSten, 2)
qcasall.glm.wls <- glm(model, gaussian, data = qcasall, n(Sten))

The first call to glm fits the model by weighted least squares, the weights being supplied by n(Sten). 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.

summary.glm(qcasall.glm.wls)
summary.glm(qcasall.glm.wls)

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

Call:  glm(formula = ChSten ~ Trt + poly(ChSten, 2), family = gaussian, data = qcasall, weights = n(Sten))

Deviance Residuals:
Min 1Q Median 3Q Max
-6.27 -1.775 -0.3197 1.708 8.19

Coefficients:  Estimate Std. Error t value
(Intercept) 1.741 0.341 5.105
Trt 0.128 0.328 0.389

Trt poly(ChSten, 2) 0.348 0.287 1.210

Residual Deviance: 43.967 on 218 degrees of freedom
Residual degrees of freedom: 1540.032 on 213 degrees of freedom.

Continued on the next page...
<table>
<thead>
<tr>
<th>Correlation of Coefficients</th>
<th>Clinic Tr.</th>
<th>pcy/18</th>
<th>pcy/2A pcy/26</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.541</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y1</td>
<td>0.018</td>
<td>-0.032</td>
<td></td>
</tr>
<tr>
<td>Trepholytigonin,211</td>
<td>0.077</td>
<td>0.117</td>
<td>-0.020</td>
</tr>
<tr>
<td>Trepholytigonin,213</td>
<td>0.153</td>
<td>0.087</td>
<td>0.136 0.088</td>
</tr>
<tr>
<td>Trepholytigonin,212</td>
<td>0.178</td>
<td>0.025</td>
<td>-0.106 0.124 0.002</td>
</tr>
<tr>
<td>Trepholytigonin,222</td>
<td>0.220</td>
<td>0.018</td>
<td>-0.240 0.003 0.206 0.0004</td>
</tr>
</tbody>
</table>

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 compute the only analyses that might be carried out on these data. Finally, even though 1990+ 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 the library for downloading via e-mail.

### ASA Biopharmaceutical Section Executive Committee—1993

**Immediate Past Chair**
Camilla A. Brooks, Ph.D.
President, CB Quaternary, Inc.
PO Box 0319
Upper Marlboro, MD 20773-0509
(301) 866-1097
(301) 866-4474 Fax

**Chair**
Bruce S. Rodda, Ph.D.
Vice President, Biosimilars & 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 Pharmacology & 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 Biometrics
Astra/Merck Site 300
300 Berwyn Park
Berwyn PA 19312
(215) 651-7879
(215) 651-7996 Fax

**Section/Financing**
John Schultz, Ph.D.
The Upjohn Company
240 Portage Road
Kalamazoo, MI 49001
(616) 385-7472
(616) 329-9348 fax

**Program Chair**
Mark Scott, Ph.D.
Director, Biometrics for Medical Systems
Clinical and Medical Affairs
ICI Pharmaceuticals
Chirico Wing 3
Boulevard 202 & 141
Wilmington, DE 19897
(302) 866-8469
(302) 866-2442 fax

**Program Chair-Elect**
Kenneth J. Koury, Ph.D.
Building 60, Room 203
Lederle Laboratories
North Middletown Road
New York, NY 10693
(516) 732-3027
(514) 733-5249 Fax

**Work Group Coordinator**
Nick Teoh, Ph.D.
BioMedics
Shering Plough Research Institute
2000 Galloping Hill Road
Raritan, NJ 07033
(908) 298-5086
(908) 298-4659 Fax

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


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: AIVITAL 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 Quantitative

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 Cnaan
Division of Biostatistics
The Children’s Hospital of Philadelphia
34th Street and Civic Center Boulevard
Philadelphia, PA 19104-4399
(215) 349-4486
(215) 996-8487 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

2679 1679 1579
DORIS L. DAVID
313 MANOR ROAD
MARLEYSVILLE PA 19458

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