Biopharmaceutical Section

American Statistical Association

Biopharmaceutical Report

Volume 4, No. 3 Fall 1996

Chair: *Gary L. Neidert* **Co-Editors:** *Curtis Wiltse and Bill Huster*

Editor's Note

Curtis Wiltse

This issue of the Biopharmaceutical Report *features several papers on adverse events. These papers, along with the two in the preceding issue, are intended to serve as a prelude to the Section-sponsored Workshop on Adverse Events, to be held October 28 and 29. Registration information is included in the August/September issue of* Amstat News. *Attendance is limited, so register early!*

An upcoming issue of the Biopharmaceutical Report *will feature reviews of books that we can recommend to our clients. Tom Bradstreet, from Merck, is organizing these reviews. Bill Huster and I welcome your suggestions for lead articles, reviews of books or software, and Section news, particularly if you volunteer to do the writing! We'd love to hear from you. In particular, we haven't had any software reviews for some time. Is there anyone out there using a new and nifty piece of software who is willing to review it for us? Please speak up.*

Have you visited the Biopharmaceutical Section's Web site yet? Sally Greenberg has done a lot of work in getting it going. The site can be accessed through the Sections page at the ASA Web site (http://www.amstat.org). Currently the site includes Section information, the Biopharmaceutical Report, *the program for the Workshop on Adverse Events, and lots more.*

Adverse Events: After 58 Years, Do We Have it Right Yet?

Joel C. Scherer and Curtis G. Wiltse

Eli Lilly and Company

1. Introduction

Safety information for drugs was first required in the United States by the Federal Food, Drug, and Cosmetic Act of 1938; hence, the pharmaceutical industry and the Food and Drug Administration (FDA) have had 58 years to resolve issues in defining, evaluating, recording, and summarizing adverse events (AEs). More recently, a description of what summary information about adverse events the FDA expects to be submitted in clinical study reports was provided in a guideline published by the FDA in 1988 (1); the International Conference on Harmonisation (ICH) Guideline (2) asks for essentially the same information. The FDA Guideline asks for:

- Frequencies of treatment-emergent events (#, N, %).
- Events grouped by body system.
- Each event divided into severity categories (if used).
- Relationship/causality (if used).
- Display original terms used by the investigator (individual study report) and group related reactions, e.g., by using a dictionary (integrated safety summary). If dictionary is not used in study report, probably synonymous reactions should be grouped.

Contents

FEATURED ARTICLES

BIOPHARMACEUTICAL SECTION NEWS

- If study size permits, more common AEs that seem to be drug related should be examined for relationship to dosage and to mg/kg dose, to dose regimen, to duration of treatment, to total dose, to demographic characteristics such as age, or to other baseline features such as renal status, and to blood level, if available. Save this analysis for the integrated safety summary if individual studies are small.
- Laboratory findings can constitute an AE (e.g., ECG abnormality suggesting infarction, serious arrhythmia, etc.).

The FDA Guideline (1) states that listings should contain:

- Age, sex, race, weight
- Treatment, dose, and mg/kg dose at time of AE
- Compliance, if available
- Date of onset, if known, or clinic visit at which event was discovered
- Duration of treatment at time of AE
- The AE, using investigators terminology
- Duration of AE
- Intensity/severity
- Action taken
- Outcome
- Relationship to test drug.

There are two different kinds of events, those generated during a clinical trial and those spontaneously reported for drugs already on the market. In the clinical trial setting, the purpose of collecting adverse event data is to enable a complete and accurate summarization of adverse events which can be expected in the population of patients that will be taking the medication, realizing that infrequent events may not be observed in the limited clinical trial patient group. The frame of reference for accuracy should be the representation of specific effects of the drug which can be used to guide the practicing physician in the use of the drug and the evaluation of events reported by a patient as to whether or not they are likely to be related to the drug.

The purpose of obtaining reports of spontaneous events is of the nature of epidemiological surveillance, to detect marked changes in frequency and seriousness of events from what was observed during clinical trials. Of particular concern are events which are serious but infrequent. The summarizing of spontaneous adverse events is often limited by the sparcity of data available about the event and the lack of a denominator to reliably gauge the relative frequency of the event of interest. In contrast, the pharmaceutical sponsor of the clinical trial can determine the amount of information about the event that they want to obtain and the number of patients exposed to the study medication is known. This paper will focus on adverse events in clinical trials.

The purpose of this paper is to identify problems related to defining and evaluating adverse events in clinical trials, and capturing data about those events. We will also make some suggestions for improvements in these areas. We will not directly address issues related to summarizing events in reports of clinical trials. We will also not address issues primarily related to the mechanics of reporting adverse events to regulatory authorities.

2. How should we define an adverse event in a clinical trial?

To our knowledge, neither the FDA nor European regulatory authorities have provided a definition of an adverse event. The exact operational definition may vary among pharmaceutical companies, but in general it encompasses the concepts of (a) any undesirable experience (b) that occurs in a clinical trial participant (c) whether or not it is considered related to the study drug, (d) even if the patient never receives study drug (intent-to-treat). In addition to varying definitions among companies, the concepts above are broad and leave room for varying interpretations.

The situation with serious adverse events (SAEs) is somewhat different. The FDA has defined that "With respect to human clinical experience, a serious adverse drug experience includes any experience that is fatal or life-threatening, is permanently disabling, requires inpatient hospitalization, or is a congenital anomaly, cancer, or overdose" (3). The ICH defines an SAE as "Any untoward medical occurrence that at any dose: results in death, is life-threatening, requires inpatient hospitalization or prolongation of existing hospitalization, results in persistent or significant disability/incapacity, or is a congenital anomaly/birth defect" (4). These definitions are more precise, which makes for more uniformity in collecting and summarizing SAEs than is the case for non-serious adverse events.

The ICH also defines other significant adverse events as "Marked haematological and other laboratory abnormalities and any adverse events that led to an intervention, including withdrawal of drug treatment, dose reduction or significant additional concomitant therapy" (2).

These definitions leave unresolved several categories of events that may be observed in clinical trials:

(1) Should expected clinical outcomes of the disease which are efficacy endpoints of the study be distinguished from SAEs?

Traditionally, all expected clinical outcomes observed during the study would be considered adverse events and reported as such. Advantages of this approach include the ability to directly observe any unexpected increased frequency of events in the treatment vs. control groups. The major disadvantage is the burden of expedited reporting of frequent clinical outcomes that are expected to be due to only the disease and not to the study drug.

Conceptually, the distinction between these outcomes and other adverse events is similar to that between solicited adverse events, where a checklist of pre-specified events is used, and spontaneously reported events. Because event rates are typically higher when events are solicited, events collected using the two methods are usually reported separately.

Recently, the FDA has permitted the separation of clinical outcomes defined as efficacy endpoints (with the exception of death) from SAEs. The clinical outcomes are not classified as SAEs and are not subject to expedited reporting. The assumption underlying this approach is that the active treatment will only decrease or leave unchanged (i.e., ineffective therapy) the frequency of observed clinical outcomes. This approach can greatly simplify and reduce the burden of expedited SAE reporting. By removing expedited reporting of a large number of expected events, a more focused assessment of underlying SAEs that may be drug related is possible.

In order to permit adequate monitoring of the safety aspect of the clinical outcomes data, they must be reported by a patient summary form in a timely fashion when each patient completes the study and be reviewed periodically by the data monitoring board.

(2) Are surgical and diagnostic procedures adverse events?

In a study of acute myocardial infarction, coronary angiography, coronary artery bypass graft surgery, placement of an intraortic balloon pump, and angioplasty are expected diagnostic and interventional procedures. Should these be recorded at all, since they are part of the expected diagnostic

and therapeutic interventions performed in this disease state? If they should be recorded, should they be recorded as adverse events, or as clinical outcomes? If these are considered to be adverse events, how should severity be assigned?

In a simpler example, if an appendectomy is performed, should appendectomy and appendicitis be recorded as adverse events, or only the appendicitis?

In evaluating this issue, we need to remember that the objective of recording and summarizing adverse events is to come to conclusions about the negative effects caused by the study drug. The drug may cause an appendicitis; it cannot cause an appendectomy, which is caused purely by a human decision. Our proposal is to record only the illness leading to the surgical or diagnostic procedure, although the procedure could be captured as part of the action taken in response to the event. Categories could include:

- surgical procedure
- medical therapy
- radiation or nuclear medicine therapy
- diagnostic procedure.

(3) What changes in laboratory variables (e.g., chemistry, hematology, urinalysis) are adverse events?

The FDA Guideline states that "Laboratory findings that constitute an adverse event (ECG abnormality suggesting infarction, serious arrhythmia, etc.) should be included" (1). Despite this, there is considerable ambiguity about if and when a laboratory finding in the absence of clinical findings should be reported as an adverse event.

In a benchmarking exercise performed by Eli Lilly and Company, it was learned that abnormal lab values are almost always entered as adverse events, although only one-third of the pharmaceutical companies surveyed definitely considered the abnormality an adverse event if there were no symptoms.

One approach is to never consider a laboratory finding alone as a primary adverse event, but rather require that the underlying clinical event be reported as the primary event; e.g., in an oliguric (decreased urine output) patient with a creatinine of 6 mg/dl, the primary adverse event would be renal failure, not elevated creatinine. The elevated creatinine would be a secondary event. Using this approach, abnormal laboratory values are reported as adverse events only when they are secondary to a primary clinical event.

The other approach would be to report abnormal laboratory values themselves as adverse events, if they fall outside a pre-specified range. The difficulty with applying this in general is the large number of abnormal laboratory findings that are often present in patients without any obvious clinical abnormality. For example, a mildly elevated creatinine in a diabetic patient. A more practical approach would be to selectively pre-specify analytes to be monitored against corresponding abnormal ranges, beyond which the result would be considered an adverse event. The analytes specified would depend on the clinical pharmacology of the drug and the clinical trial setting.

3. How should we evaluate adverse events in clinical trials and capture data about those events?

(1) Is there a better coding system than COSTART for assigning class terms?

The FDA Guideline recommends the use of a dictionary for grouping similar events, but adds that "Experience at this time is too limited to recommend a particular one for this purpose" (1), although the COSTART dictionary is widely used. Potential problems with any dictionary include mapping synonymous clinical events to different class terms (the one-to-many problem), or mapping clinically different events to a single class term (the many-to-one problem). An example of the first problem is that the events of heart attack, heart arrest, myocardial infarction, and coronary attack, which all mean cardiac arrest (which itself has no COSTART class term), map to the class term of cardiovascular disorder, asystole, myocardial infarct, and coronary artery disorder, respectively. An example of the second problem is that angiography, CABG surgery, coronary angioplasty, and intraortic balloon pump all map to surgical procedure. Another example is that cardiomyopathy, left ventricular aneurysm and thrombus, heart murmur, permanent pacemaker, rheumatic heart disease, heart size change, and heart attack, along with approximately seventy other events, all map to the single COSTART class term of cardiomyopathy. Another problem which causes confusion is non-descriptive class terms such as reaction unevaluable.

Avoiding these mapping problems involving translation of clinical terms to standardized class terms in a dictionary is difficult. However, they can potentially be reduced or eliminated by use of a more appropriate dictionary, especially if classifications were consistent with the ICD9 codes already used by health care professionals. The MEDDRA dictionary (5), which has been adopted by the ICH, is one example of an effort in this direction.

(2) How should syndromes be recorded on the case report form? As the syndrome, as each symptom separately, or both?

A syndrome is a collection of signs, symptoms, or laboratory findings which together describe a distinct clinical entity. For example, the syndrome of congestive heart failure is a pathophysiologic entity described by one or more of the following: shortness of breath, peripheral edema, easy fatigueability, pulmonary rales, elevated venous pressures, low plasma sodium, elevated BUN, etc. Other examples include pneumonia or other infections with systemic manifestations (e.g., influenza).

An issue in the consideration of recording syndromes or symptoms is how well-defined a clinical syndrome is, in terms of clinical criteria that can be used to make the diagnosis. Not all syndromes are well-defined. Some, such as congestive heart failure, are well-defined, whereas others, such as pneumonia, are not.

Summarizing both the syndrome and its components presents two problems, each of which may misrepresent the effect of the drug. A general problem is that the symptoms which occur alone are not distinguished from those that occur as part of a syndrome. A more specific problem is that the increased incidence of a true adverse event related to the drug may be lost in the high incidence of the same event which occurs as a component of the syndrome in both groups.

Nickas gives an example of intracranial hypertension, one manifestation of which is headache (6). If three of 100 (3%) patients in a clinical trial develop intracranial hypertension with associated headache after randomization to study drug, and an additional five (5%) patients develop headache of unknown etiology, what should be reported as the incidence of treatment-emergent headache? Nickas suggests resolving this problem by recording the syndrome, if it is known; the event itself, such as headache, should be recorded and reported as the primary adverse event only if the syndrome is unknown. If the syndrome is known, the components or associated outcomes could be recorded as secondary adverse events.

(3) How should primary vs. secondary adverse events be recorded?

A related issue with many of the same summarization problems is an adverse event (the primary event), such as bleeding, with consequences (secondary events), such as

hypotension, oliguria, altered mental status, weakness, secondary angina, or heart failure, etc., which are themselves adverse events. Problems for both this and the previous issue include multiple counting of the same events and the over-reporting of components of a syndrome or secondary events, which are not in and of themselves related to the drug. These lead to inaccurate conclusions about the undesirable effects of the drug. It is the analysis of the primary adverse events which leads to correct understanding of the causal relationships between drug administration and the occurrence of adverse events. Both primary and secondary adverse events can be recorded on the CRF, with the secondary events clearly linked to the primary event or syndrome.

(4) What kind of algorithm should be used for assessing the relationship between the adverse event and the study drug (causality)?

One approach is to not assign causality for individual events, but use the comparison of frequencies between the study drug and control treatment groups to assess causality. This method assesses causality for groups of patients rather than for individual patients. It is also a retrospective method.

Prospective assessment of causality for individual events is a regulatory requirement for serious adverse events. Causality could be defined for individual events using a pre-specified algorithm containing information regarding the known pharmacology of the drug, preclinical observations, relationship to drug administration, and the patient's history and characteristics. However, at the present time, the lack of a standardized approach to developing algorithms for assigning causality in a clinical trial setting results in variable assignment of causality among reporters.

(5) Since defining severity in terms of functional impact doesn't always make sense, should the definition be expanded to include other effects, such as physiologic ones?

The FDA Guideline requirements for describing an adverse event ask for a description of intensity (e.g. mild, moderate, severe) (1), which is usually defined in terms of its functional impact on the patient, i.e., interference with usual daily activities. For certain events, this functional description may be inadequate. For example, bleeding without physiologic impact on the patient but requiring blood transfusion may be considered mild by these criteria, although the need for a transfusion may indicate that a severity of moderate or severe may be more appropriate, especially considering that the transfusion is masking symptoms that might otherwise be observed. Thus, the true severity of the bleeding is underrated. An alternative scale of severity could be defined, based on the clinical pharmacology of the drug, that would more accurately reflect the true severity of the event. This scale could be based on not only functional impact, but also physiological impact, need to intervene with medical or other therapy, patient discomfort, and health risk to the patient.

For important events, a severity scale could be defined specifically for that event. For example, the GUSTO trial, which compared four thrombolytic strategies for acute myocardial infarction, defined bleeding complications as severe or life-threatening if they were intracerebral or if they resulted in substantial hemodynamic compromise requiring treatment. Moderate bleeding was defined by the need for transfusion. Minor bleeding referred to other bleeding, not requiring transfusion or causing hemodynamic compromise (7).

Another approach, rather than defining severity precisely, would be to consider whether even the best-defined severity scales really tell us much, since a more severe manifestation of an event becomes an event with a different name. For example, as dizziness becomes more severe, it becomes syncope.

(6) How should adverse events be collected in a short-term closely-monitored trial (e.g., a 48-hour inpatient infusion in myocardial infarction patients) as compared with a long-term outpatient trial (e.g., a 6-month trial with monthly visits in patients with asymptomatic hypertension)?

Medical treatment for a myocardial infarction may include a 48-hour infusion of study drug, under close medical supervision in a hospital. The intense observation of the patient in this setting leads to the potential for super-reporting of adverse events. An adverse event that may be reported by an asymptomatic hypertensive patient in an outpatient setting as being of one day's duration may be recorded as 6 intermittent events for the myocardial infarction patient. For each episode, the severity, relationship to study drug, outcome, etc. may be recorded, whereas for the hypertensive patient, there may be only one record with this information. Also, the intensity of observation by the study nurse and other personnel may result in AEs being recorded which would not be recorded in an outpatient setting.

What are the important features of the plenitude of AE data from this setting that should be extracted and summarized in a clinical study report? Is it important that an event had six episodes during one day? Is it important to report events that would not even be recorded in an outpatient setting?

Another issue is how intermittent events should be recorded and reported. Do the start and stop dates and times always need to be recorded for each episode, or are there events for which it is adequate to record only that the event occurred intermittently over a certain time interval, as well as the highest severity during that interval? Does it matter how long that interval is; i.e., should more (or less) information be recorded if the event occurs intermittently over one day as compared with over one month?

How should the frequency of treatment emergent events be reported? Once per patient, the number of occurrences per patient, or the duration of events per patient? If once per patient, a patient may have many occurrences per day in a continuously monitored setting. So if each patient is counted only once, why collect so much data? The amount of detailed information that is collected needs to be balanced against what is needed to achieve the objective of accurately summarizing safety information.

On the summarization side, should the definition of treatment-emergent be different for intermittent events?

(7) For events which do not meet the regulatory definition of serious, should the amount of information that is collected be related to the seriousness and relatedness of the event?

Should as much information be recorded for a garden-variety case of the sniffles as for a cardiac arrhythmia? Realizing that some adverse events are deemed to be causally related to the drug only when examined under the retrospectoscope, is it possible to define conditions under which less-than-full information can be recorded for an event? Should less information be collected for events which are not considered by the investigator to be at least possibly causally related to the drug, at least in the setting of a Phase III trial? If the common adverse events for a drug have been well characterized in Phase II, there may be circumstances under which it is appropriate to consider discussing with the regulatory authorities the recording of only SAEs, or only events which are serious and at least possibly causally related to the drug.

4. Summary

We have outlined some of the problems encountered in defining and evaluating adverse events in clinical trials, and

capturing data about those events. These problems arise, in part, due to a need to balance the collection of adequate data to describe adverse events with what is reasonable and practical to collect in a clinical trial setting. The reason for collecting exhaustive data is that for the vast majority of events we are trying to use a comparison between treatment groups of (a) relative frequencies of events and (b) characteristics of those events to make up for not being able to assess causality with certainty for an event in an individual patient. This epidemiological approach is based on the accuracy of the event description and the resulting frequencies reported, and we have pointed out some of the difficulties in accomplishing this using current approaches. In addition, improving the accuracy of description will have the added benefit of providing the most useful descriptive clinical information for the treating physician.

Because of the complexity of the problems involved, there is no single correct way to solve these problems; instead, they may be best resolved by worldwide agreement by regulatory agencies and sponsors on a standardized approach to defining, evaluating, recording, and summarizing adverse events.

5. References

1. Center for Drug Evaluation and Research, U.S. Department of Health and Human Services, Public Health Service, Food and Drug Administration (1988), *Guideline for the Format and Content of the Clinical and Statistical Sections of New Drug Applications*. July, pp. 71-74.
2. ICH Expert Working Group (1995), *ICH Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH Harmonised Tripartite Guideline: Structure and Content of Clinical Study Reports, Section12*, November 30.
3. Department of Health and Human Services, Food and Drug Administration (1994). *Code of Federal Regulations, 21 CFR 312.32(a) IND Safety Reports*, Revised April 1, pp. 82.
4. *ICH Guideline for Industry: Clinical Safety Data Management: Definitions, and Standards for Expedited Reporting.* Quoted in: *ICH Expert Working Group. ICH Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH Harmonised Tripartite Guideline: Structure and Content of Clinical Study Reports, Section 12.* November 30, 1995.
5. Wood, K. L. (1994), "The Medical Dictionary for Drug Regulatory Affairs (MEDDRA) Project," *Pharmacoepidemiology and Drug Safety*, 3:7-13.
6. Nickas, J. (1995), "Adverse Event Data Collection and Reporting: A Discussion of Two Grey Areas," *Drug Information Journal*, 29:1247-1251.
7. Topol, E. J. for the GUSTO Investigators (1993), "An International Randomized Trial Comparing Four Thrombolytic Strategies for Acute Myocardial Infarction" *N Engl J Med*, 329:673-682.

A Statistical Perspective on Adverse Event Reporting in Clinical Trials

Janet Wittes

Statistics Collaborative

Abstract

Although a central objective of many randomized clinical trials is assessment of the safety of the study treatment, statistical input into the process of reporting on safety and of interpreting the data on safety is often relegated to very low priority. Unfortunately, the task of accurately characterizing the safety profile of a treatment under study is difficult both during the interim monitoring of the trial and at the end of the study. The data tend to be messy and idiosyncratically collected. Problems of multiplicity, low sample size, unspecified prior hypotheses, and unhelpful classification add complexity to the task of identifying reliable associations between treatment and adverse events. Finally, many investigators fail to compare randomized treatment groups in assessing adverse events within the context of a randomized clinical trial. This paper sketches some reasons for the difficulties in analyzing safety data and proposes some simple approaches to enhancing the reporting.

Keywords: adverse events, clinical trials, data monitoring, safety data

1. Introduction

Although a central objective of many clinical trials is assessment of the safety of the study treatment, statistical input into the process of reporting on safety and of interpreting the data on safety is often relegated to very low priority. Unfortunately, the task of accurately characterizing the safety profile of a treatment under study is difficult. Nonetheless, this paper argues that if we are truly serious about describing the adverse effects of a therapy, analyses of safety should not be merely tag-a-longs to our analyses of efficacy.

A Data and Safety Monitoring Board (DSMB) typically considers safety during the process of interim monitoring of the study and the investigator or sponsor summarizes the information at the end of the study. This paper addresses some statistical problems confronting the DSMB in its safety monitoring role and the data analyst in the final assessment of safety. One role of both interim and final assessment is flagging the possibility of an association of "serious and unexpected" adverse events with the therapy under study. The tools to perform this task are blunt indeed. A plethora of tables and graphs that describe safety may bury some true signal in a cacophony of numbers. Although an important statistical task is to enhance that signal, I suspect I am not the only statistician who tunes out when a DSMB begins to discuss the occurrence of adverse events, for we statisticians often find the multiplicity inherent in the reporting of adverse events so overwhelming that we tend to discount what may be real phenomena. We hope that by the next meeting of the DSMB some Law of Large Numbers will have all but erased the differences currently observed. Many clinicians, on the other hand, confronted with such data, seem to sift through the tables with a fine tooth comb to detect "clinically significant" effects of a treatment.

This paper describes some problems faced in discerning the degree to which a therapy being tested is giving rise to adverse experiences. I outline some of the common methods for reporting safety data, suggest strategies for enhancing those methods, and pose some questions for the future. Although I focus primarily on the reporting of events during an ongoing clinical trial, many of the same problems confront the interpretation of data at the end of the study. I shall use the terms "adverse event," "adverse experience," and "safety data" interchangeably to refer loosely to "bad things that happen." Certain fields, for example cancer therapeutic trials, also refer to "toxicity data" or "toxicities"; "adverse drug reaction" is another common term in the context of the study of drugs. Throughout, this paper refers to "treatment" and "control"

groups, but the considerations pertain as well to an arbitrary number of study groups.

The process of safety monitoring asks at least two distinct types of questions. First, for a type of adverse event already known to be related to study medication, it monitors the incidence and severity of the event, in part to ensure that participants in the trial are not experiencing an excessive level of risk. Second, the safety monitoring process looks for evidence that some types of adverse event, as yet unknown, are related to a treatment under study. Estimation of the frequency of a known type of event is conceptually simple, but often, for reasons described below, fraught with practical problems. Much more difficult, both conceptually and practically, is the task of associating an as yet unidentified type of adverse experience with a treatment under study. Many events can and will occur to participants in the typical clinical trial. If the study is a so-called treatment trial in a serious disease, the ordinary development of disease will entail a host of related adverse, often serious, experiences. If, on the other hand, the study investigates a strategy for prevention of disease in a healthy population, the trial may continue for several years and diverse adverse experiences will occur to people as part of the natural course of their lives. Thus, it is not an easy task to deduce whether a treatment under study increases the probability of a specific type of adverse experience or whether the study treatment caused a specific individual adverse event.

2. Hypothesis-free Searching, Data Quality, and Power

In testing the effect of a therapy on the primary endpoint of a clinical trial, the well-trained trialist knows to begin with a clear hypothesis. That hypothesis drives the formal definition of the endpoint as well as the sample size of the study. The investigators spend considerable time and effort during the design and implementation of the trial to measure the primary endpoint reliably. Planning committees spend many hours reducing the number of primary endpoints to as close to one as possible in order to design a trial with adequate power to answer the most important questions. Case definition for the disease under study and for the primary endpoint may require such measures as confirmatory tests, endpoint adjudication committees, and blinded readings to ensure that an event declared an endpoint truly corresponds to an endpoint. In the language of diagnostic screening, the design of a typical clinical trial emphasizes the need for high specificity, or a low false positive rate, in the identification of endpoints.

Many trials also institute strategies for active case ascertainment. Such requirements as frequent visits, telephone follow-up, or repeated laboratory tests all increase sensitivity, that is, the likelihood of eliciting an event that occurs.

Even with this concerted attention to the parsimony of questions asked, the intensity of endpoint ascertainment, and the reliability of its measurement, clinical trials often have low power to identify the effects they were designed to detect. Very few trials declare as an objective the estimation of the event rate for the primary outcome measure, for experienced trialists are well aware that in most trials the event rate is different from the rate in the general population (1).

In contrast to the care with which the protocol of the trial specifies the endpoints pertinent to efficacy, the case definition of most adverse events is loose, the method of ascertainment nonspecific, and the typical sample size grossly inadequate to identify differences between treated and control groups for all but the most frequent types of events. It should not be surprising, therefore, that the usual clinical trial has low power for determining whether many specific adverse experiences are associated with the test therapy. On the other hand, while trialists are justifiably reluctant to apply the event rate of the primary efficacy outcome in the trial to the population at large, many people assume that the event rates of adverse experiences observed in a trial are applicable to a larger population. Worse yet, trialists sometimes lull themselves into believing randomization is unnecessary for identifying unusual adverse events, for they argue that they can safely assume the occurrence of an event that is ordinarily rare in the disease studied is a consequence of the study treatment. Investigators often appeal to this argument to avoid entering placebo patients in Phase I studies. For example, in a study of a recombinant human ciliary neurotrophic factor (rHCNTF) in patients with amyotrophic lateral sclerosis (ALS, or, Lou Gehrig's disease), a "seemingly unusual incidence of [mouth sores] was observed" (2). Since this condition is atypical in ALS patients, a finding of nine cases in 43 study participants (21%) would have led to a strong suspicion that the study medication had caused the problem. Fortunately, this Phase I study had included a placebo group; since three of the 14 placebo patients (21%) also complained of mouth sores, the finding was attributed not to the adverse effects of rHCNTF, but to an unusual type of patient or perhaps a very detailed elicitation of symptoms.

A clinical trial necessarily attempts to associate adverse experiences with a study treatment in the context of no prior hypotheses for unexpected findings. After all, a purpose of the trial is to detect events in the absence of prior information or suspicion concerning association. In addition, the quality of the data collected tends to be low, largely because no study has the resources to collect accurate data on every possible event. The question faced, then, is how to make sensible conclusions from incomplete and poorly classified data, low power, and daunting multiplicity.

3. Adverse event collection and enhanced reporting of safety

Adverse event reporting typically includes at least four different types of data collection, each with its own set of statistical issues: laboratory data and parameters measuring vital signs; data concerning the occurrence of events that are known to be related to the study treatment; a list of all adverse events regardless of severity or likely association with study treatment; and an ongoing collection of serious and unexpected events reported to the sponsor and perhaps the DSMB as soon as possible after they occur. When the statistical charge is to prepare or interpret these types of data, we should elicit from our clinical colleagues information on what is already known, what is suspected, and what is unknown. Our reporting should clearly distinguish these types of information.

The data analyst must acknowledge that information on adverse events collected in a randomized clinical trial is, like information on efficacy, the result of an experiment. The analyst must not discard the rules for unbiased comparisons in the study of adverse events. I believe that the primary determination of safety, like the primary determination of efficacy, should analyze the participants within the groups to which they were randomized, not the according to the treatment they received. This standard requires, admittedly, a hard sell to clinicians who will argue strenuously for an as-treated analysis. Such an approach does not preclude looking at the adverse events among those who received their randomized assignment, but interpretation of both the as-randomized and as-treated groups should recognize all the usual cautions about inference from the respective analyses.

Many times we statisticians present data on safety without clear thought to making the tables interpretable. We often blame the FDA for our failure to present data in a sensible way by asserting, "The FDA insists on presenting 3 digits for p-values," or "The FDA requires that we list the adverse events in

Table 1. Percentiles for laboratory parameters and vital signs

		Control Group					Treatment Group				
Parameter		10	25	50	75	90	10	25	50	75	90
Heart rate -	observed	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx
	expected	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx
BUN -	observed	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx
	expected	xx	xx	xx	xx	xx	xx	xx	xx	xx	xx

descending order by frequency of WHOART code." Where is it written? My FDA friends, both the statistical and clinical reviewers, look at me askance when I ask them, "Do you really require that we...?" The answer is usually something like, "I don't know why the drug companies think they have to send us ...." In reporting to DSMBs, the FDA, and the scientific community at large, our collective goal ought to be to present data sensibly in a way that eases interpretation. (If your company insists that the FDA demands something that seems difficult to interpret, then prepare that too, but don't neglect to present what makes sense.)

3.1 Laboratory data and vital signs

The collection and review of clinical laboratory data and vital signs provide clinical or subclinical indications of potential problems caused by treatment. If the monitoring of laboratory data shows elevations in the rate in which dangerous laboratory values occur, then the data provide some warning about the potential toxic effect of the therapy. This information may lead to redoubling of efforts to look for specific clinical signs or symptoms. Similarly, a difference in the distribution of vital signs in the treatment group may indicate a problem with one treatment arm. While such data often show very large differences in laboratory parameters between the treatment groups with unambiguous statistical significance, the monitoring committee may not know how to act. The study therapy is supposed to be active, so changes in laboratory parameters may simply reflect the known, or perhaps unknown, metabolic pathway of the therapy. The challenge for monitoring is to predict whether the changes are so large and so frequent as to cause alarm or to lead to a recommendation to change the dose.

The frequency of reporting of laboratory and vital signs data ordinarily coincides with the regular meetings of the DSMB. In my experience, DSMBs often spend little time reviewing laboratory data, in part because such data are hard to interpret even when they are correct, but in part because laboratory data are often very messy. The database may consist of numerical data entered as character variables riddled with signs like ">" or "<" or "~," notations like "n.a." or "-," or units of measurements included in the data field. Thus, while abnormal laboratory data should serve as sentinels to the DSMB for upcoming adverse events, often the data are so dirty that they cannot be interpreted until long after they are collected.

At the end of the study, the data analyst can usually categorize quite well the differences between treated group and control with respect to these parameters, unless many of the participants in the study have died or dropped out before the data were collected.

For laboratory data and vital signs, the typical set of tables simply records means, or perhaps medians, and some measure of variability. More useful are those presentations that show other percentiles either in tabular form or graphically. Of particular interest are the 10th and 90th or the 5th and 95th percentile. Maxima and minima, while commonly presented, often serve more to diagnose the quality of the data than to provide a picture of the distribution of the values in the study group. For example, a minimum of zero often indicates improper transfer of missing values from one database to another, while a huge maximum may reflect decimal errors or failure to convert from one scale of measurement to another. For this type of data, I like to see a table with both observed and expected percentiles where the expected values are elicited from the clinical investigators. (See, for example, Table 1.) Safety does not mean no difference between placebo and treated; it means unacceptably large changes.

Tables or graphs of changes from baseline or plots over time are often quite helpful in assessing whether changes are transient or long-term. A table showing the frequency of values beyond some "alert" level is another informative presentation of laboratory data and vital signs. Tables uncluttered with too many significant digits are easier to read than tables exuding false precision.

3.2 Already known adverse events

A second type of data is collected in a trial studying a well-characterized treatment procedure for a new indication or in an as yet uncharacterized population. Here, data are collected to capture the frequency of the known adverse experiences in order to estimate the difference in the frequency of a defined event in the treated and control group. For example, the investigators of the Postmenopausal Estrogen/Progestin Interventions (PEPI) Trial, aware that the use of unopposed estrogen leads to an increase risk of uterine cancer, required annual endometrial biopsy for all participants (3). Collecting information on known adverse experiences allows assessment of the costs and benefits of a therapy. If the trialists exercise care in the definition of the adverse events and in the collection of the supporting data, this type of adverse event monitoring is reasonably straightforward. If the expected benefit in the specific trial does not offset the frequency and intensity of the adverse experiences, the trial may be halted.

Under ordinary circumstances, these types of data are reported to the DSMB at its usual meetings. Because the questions are focused, the presentation can be simple and clear. Often, however, the tables of adverse events fail to distinguish those already known from those that are unknown or those with active elicitation or symptoms from those with spontaneous reporting. Two sets of tables, one for the events that are unexpected and one for those that had been expected, are helpful. To assess the frequency of known adverse events, trials ought to institute operational definitions of those events and the protocols should include power calculations that describe the likelihood of characterizing the rates precisely. As noted above, however, a study group in a trial differs from the population at large, so that the adverse event rate may differ as well. Further, the method of eliciting events may increase the reporting of them.

3.3 All adverse events

As the third part of safety monitoring, the DSMB receives a list of all events, no matter where they fall along the spectrum of severity and no matter how likely the investigator deems their relationship to study therapy. Typically, the DSMB receives such data not only at its regular face-to-face meetings, but at intervals between those meetings as well. I find much of this information quite confusing both to read when I am a member of a DSMB and to prepare when I am the statistician

responsible for presenting the data. Trials may have many more adverse events than participants, so that the chance of seeing something that looks surprising is very high. To simplify matters and focus thinking on the important questions, at least one table is usually presented listing only the adverse events the investigator considers at least possibly related to study treatment. While such a classification eliminates some obviously unrelated events, this type of table, like the more inclusive presentations, is hard to interpret. The thresholds for declaring an event "related" to study treatment differ from investigator to investigator. Even more important, I suspect that delayed effects of a drug, even if real, will be less likely to be reported as "related" than immediate effects even if the event is not related to study treatment.

The typical listing shows all data in descending frequency by COSTART or WHOART code, often in all capital letters because the data come from the computer in that form. Too many decimal points and too many percentages clutter the table. Adverse events with the same frequency are listed alphabetically. Thus, random occurrences of events and lexicographical accidents jointly conspire to produce a table that is very difficult to read. Table 2 presents a typical example of such a presentation.

How should one respond to such a table? The eye, trained to find little p-values, alights on the 0.005, which corresponds to 7 cases of bronchitis in the treated group and none in control. The brain tries to pull together similar events (e.g., vertigo and dizziness, or all the arrhythmias), but they are far from each other in the table.

Some very simple changes in format lead to a dramatic

Table 2. Number and percent of WHOART-coded adverse events by treatment (T) and control (C) group

	T (n=142)		C (n=154)		Total (n=296)		
WHOART Code	n	%	n	%	n	%	p-value
Cardiac failure	9	(6.3%)	6	(3.9%)	15	(5.1%)	0.339
Chest pain	4	(2.8%)	8	(5.2%)	12	(4.1%)	0.300
Angina pectoris	5	(3.5%)	6	(3.9%)	11	(3.7%)	0.865
Cardiac failure left	8	(5.6%)	2	(1.3%)	10	(3.4%)	0.039
Renal failure acute	2	(1.4%)	7	(4.5%)	9	(3.0%)	0.116
Unstable angina	6	(4.2%)	3	(1.9%)	9	(3.0%)	0.254
Myocardial infraction	6	(4.2%)	2	(1.3%)	8	(2.7%)	0.121
Tachycardia ventricular	4	(2.8%)	4	(2.6%)	8	(2.7%)	0.907
Abdominal pain	3	(2.1%)	4	(2.6%)	7	(2.4%)	0.784
Bronchitis	7	(4.9%)	0	(0.0%)	7	(2.4%)	0.005
Cardiac arrest	2	(1.4%)	5	(3.2%)	7	(2.4%)	0.298
Arthralgia	0	(0.0%)	6	(3.9%)	6	(2.0%)	0.017
Cerebrovascular disorder	2	(1.4%)	4	(2.6%)	6	(2.0%)	0.468
Diabetes mellitus aggravated	2	(1.4%)	4	(2.6%)	6	(2.0%)	0.468
Hypotension	4	(2.8%)	2	(1.3%)	6	(2.0%)	0.354
Infection	4	(2.8%)	2	(1.3%)	6	(2.0%)	0.354
Peripheral Ischemia	3	(2.1%)	3	(1.9%)	6	(2.0%)	0.920
Anemia	4	(2.8%)	1	(0.6%)	5	(1.7%)	0.148
Diarrhea	4	(2.8%)	1	(0.6%)	5	(1.7%)	0.148
Dyspnea	4	(2.8%)	1	(0.6%)	5	(1.7%)	0.148
Fracture accidental	0	(0.0%)	5	(3.2%)	5	(1.7%)	0.030
Pneumonia lobar	3	(2.1%)	2	(1.3%)	5	(1.7%)	0.587
Urinary tract infection	1	(0.7%)	4	(2.6%)	5	(1.7%)	0.207
Injury - accidental	1	(0.7%)	3	(1.9%)	4	(1.4%)	0.354
Toxicity due to non-study drug	1	(0.7%)	3	(1.9%)	4	(1.4%)	0.354
Cardiac failure right	3	(2.1%)	0	(0.0%)	3	(1.0%)	0.070
Infection Bacterial	2	(1.4%)	1	(0.6%)	3	(1.0%)	0.515
Pulmonary edema	2	(1.4%)	1	(0.6%)	3	(1.0%)	0.515
Sudden death	3	(2.1%)	0	(0.0%)	3	(1.0%)	0.070
Angina pectoris aggravated	0	(0.0%)	2	(1.3%)	2	(0.7%)	0.173
Anorexia	1	(0.7%)	1	(0.6%)	2	(0.7%)	0.954
ArrYhthmia atrial	1	(0.7%)	1	(0.6%)	2	(0.7%)	0.954
Fever	1	(0.7%)	1	(0.6%)	2	(0.7%)	0.954
GI neoplasm malignant	0	(0.0%)	2	(1.3%)	2	(0.7%)	0.173
Headache	2	(1.4%)	0	(0.0%)	2	(0.7%)	0.139
Hypotension postural	1	(0.7%)	1	(0.6%)	2	(0.7%)	0.954
Malaise	0	(0.0%)	2	(1.3%)	2	(0.7%)	0.173
Suicide attempt	2	(1.4%)	0	(0.0%)	2	(0.7%)	0.139
Thrombophlebitis deep	2	(1.4%)	0	(0.0%)	2	(0.7%)	0.139
Vertigo	2	(1.4%)	0	(0.0%)	2	(0.7%)	0.139
Vomiting	2	(1.4%)	0	(0.0%)	2	(0.7%)	0.139
Abcess	1	(0.7%)	0	(0.0%)	1	(0.3%)	0.297
ArrhYthmia	1	(0.7%)	0	(0.0%)	1	(0.3%)	0.297
Arteriosclerosis	0	(0.0%)	1	(0.6%)	1	(0.3%)	0.336
Ataxia	0	(0.0%)	1	(0.6%)	1	(0.3%)	0.336
Back pain	0	(0.0%)	1	(0.6%)	1	(0.3%)	0.336

improvement in presentation. Reporting only counts without percentages, using upper and lower cases, and sorting by body system gives a much clearer picture of the distribution of adverse experiences (Table 3).

Even this table is not sufficiently clear. The alphabetical sorting puts certain similar diagnoses close to each other (e.g., all the cardiac failures are now in consecutive order), but others still are separated (e.g., "pain" is far from "back pain"). Moreover, the standard classification may not be relevant to the specific study. For example, although classifying "sudden death" or "chest pain" in the "body as a whole" category makes sense for some purposes, this study clearly includes participants at high risk for cardiovascular events. Therefore, both "sudden death" and "chest pain" are likely to be arrhythmic events and should be classified as a cardiac disorder, not assigned to "body as a whole." The error incurred by lumping these nonspecific events with the likely body system is probably less important than the error incurred by splitting them away from the rest of the arrhythmic events. Of course, redefining classifications invites data-driven results. One very useful approach for avoiding data-driven results is to classify the data on the basis of the diagnosis and total column alone (Table 4).

Table 3. Number of adverse events by treatment group and body system

Body system (WHOART Code)	T	C	Total
Autonomic nervous system disorders			
Hypotension postural	1	1	2
Hypotension	4	2	6
Body as a whole—general disorders			
Abcess	1	0	1
Ascities	1	0	1
Back pain	0	1	1
Chest pain	4	8	12
Edema	0	1	1
Fever	1	1	2
Infection	4	2	6
Infection Bacterial	2	1	3
Injury - accidental	1	3	4
Malaise	0	2	2
Moniliasis	1	0	1
Necrosis ischemic	0	1	1
Pain	0	1	1
Previously scheduled surgery	2	2	4
Sepsis	0	1	1
Sudden death	3	0	3
Toxicity due to non-study drug	1	3	4
Cardio-vascular disorders, general			
Cardiac failure	9	6	15
Cardiac failure left	8	2	10
Cardiac failure right	3	0	3
Circulatory failure	1	0	1
Unstable angina	6	3	9
Central and peripheral nervous system disorders			
Ataxia	0	1	1
Dizziness	1	0	1
Endocrine disorders			
Headache	2	0	2
Vertigo	2	0	2
Tremor	0	1	1
Gastro-intestinal system disorders			
Abdominal pain	3	4	7
Diarrhea	4	1	5
Gastric ulcer	1	0	1
Gastric ulcer hemorrhagic	1	0	1
Gastritis	1	0	1
GI hemorrhage	2	2	4
Hemorrhage rectum	0	1	1
Vomiting	2	0	2
Heart rate and rhythm disorders			
Arrhythmia	1	0	1
Arrhythmia nodal	1	0	1
.	.	.	.
.	.	.	.

Table 4. Adverse events by treatment (T) and control (C) group and category of event.

	T	C	Total	T/C
Angina/chest pain/MI	15	19	34	0.8
Heart failure	20	8	28	2.5
Arrhythmia	10	14	24	0.7
Cardiac arrest/sudden death	5	5	10	1.0
Bleeding	6	4	10	1.5
Headache/dizziness	5	0	5	5/0
Infection	8	4	12	2.0
Injury	1	8	9	0.1
Carcinoma	0	4	4	0/4

Note: T/C is the ratio of the number of events in the treated to the number of events in the control groups. If either group has no events, the ratio is presented as 0/x or x/0.

3.4 Serious and unexpected events

The last type of reporting, though very important, also leads to difficulties in interpretation. The FDA requires serious and unexpected events to be reported separately, even if they bear no plausible relationship to the study treatment.

All events that are either serious, or unexpected even if not serious, are reported to the Data and Safety Monitoring Board in as close to real time as possible. The DSMB is asked to make sense of this sporadic, often incomplete, set of reports. The data come unaudited. Sometimes it is not clear that the rate of reporting is the same in both study arms. A good denominator is very hard to find, so the calculation of event rates is problematic. In order to help the DSMB interpret the data, I find it useful to see tables of "dirty" data, which will include errors of all sorts and ill-defined denominators, "clean" data, which because it summarizes only audited data will include a smaller sample size, and a table that shows what proportion of dirty data are likely to be misclassified.

4. Other issues in the statistical assessment of safety

Although this paper has discussed safety without reference to efficacy, the two must in fact be viewed together, for only in the context of an efficacious therapy are adverse experiences tolerable. Canner (4) has discussed considerations related to monitoring data for evidence of both safety and efficacy. Some authors recommend constructing scenarios for monitoring both safety and efficacy (5). While conceptually such combination is very attractive, the methods should be used with caution. As mentioned above, in a typical trial, data on safety arrive much earlier than data on efficacy. For example, in a trial studying the long-term effects of a new therapy for diabetes, the adverse experiences will emerge long before the data can show decreases in the probability of occurrence of the long-term sequelae of the disease. Similarly, cancer chemotherapeutic agents will declare their toxicity before data have accrued that can address the benefit of treatment.

At the end of the trial, a balanced judgment must be made, either formally or informally, to assess the net effect of treatment. Chuang-Stein et al. (6) have proposed methods for combining information on safety and efficacy in order to judge the relative benefits and risks of a therapy.

In many studies, especially those with mortality as part of the primary endpoint, safety and efficacy are intertwined. Thus monitoring for safety also monitors for efficacy (7). Although many people have argued that the process of statistical data monitoring can separate safety from efficacy, I find the reasoning unconvincing. I believe we must develop α-sparing statistical methods for data monitoring that acknowledge that when a DSMB looks at safety data, it is also looking at efficacy data. Often that will mean the DSMB must look at the data unmasked to treatment group.

Formal analysis of safety data compares the event rates in the treated and control groups. Thus, if in a trial we see 5 events of a single type, we ask whether the distribution of those 5 events into treated and control could reasonably have happened by chance. Because the total number of events is low, the conditional probability of any allocation is not very small. For rare events, however, the important, though often unanswerable, question is the unconditional one: is the number observed in the treatment group surprisingly large?

Safety data are challenging for the statistician. Our usual methods of inference seem ineffective for addressing the questions we would like to ask. Statistical testing is mired in multiplicity; confidence intervals are too wide to be of use; appeal to a Bayesian framework does not, I believe, help. We should strive to present the data in a sensible way. We must not allow fear of squandering our α to prevent us from monitoring safety as carefully as we know how.

References

1. Steering Committee of the Physicians' Health Study Research Group (1989), "Final Report on the Aspirin Component of the Ongoing Physicians' Health Study," *N Engl J Med.*, 321: 129-135.
2. The ALS CNTF Treatment Study (ACTS) Phase I-II Study Group (1995), "A Phase I Study of Recombinant Human Ciliary Neurotrophic Factor (rHCNTF) in Patients with Amyotrophic Lateral Sclerosis," *Clin Neuropharm*, 18: 515-532.
3. Espeland, M. A., Bush, T. L., Mebane-Sims, I., Stefanick, M. L., Johnson, S., Sherwin, R., and Waclawiw, M. (1995), "Rationale, Design, and Conduct of the PEPI Trial," *Con Clin Trials*, 16: 3S-19S.
4. Canner, P. (1983), "Monitoring of the Data for Evidence of Adverse or Beneficial Treatment Effects," *Cont Clin Trials*, 4: 467-483.
5. Freedman, L., Anderson, G., Kipnis, V., Prentice, R., Wang, C. Y., Rossouw, J., Wittes, J., and DeMets, D. (1996), "Approaches to Monitoring the Results of Long-Term Disease Prevention Trials: Examples From the Women's Health Initiative," *Cont Clin Trials*, 17: xxx-xxx.
6. Chuang-Stein, C., Mohberg, N. R., and Sinkula, M. S. (1991), "Three Measures for Simultaneously Evaluating Benefits and Risks Using Categorical Data From Clinical Trials," *Stat in Med*, 10: 1349-1359.
7. PMA Biostatistics and Medical Ad Hoc Committee on Interim Analysis (1993), "Interim Analysis in the Pharmaceutical Industry," *Cont Clin Trials*, 14: 160-173.

How Blind are Double-Blind Studies When the Product Exhibits a Very Distinct Safety Profile?

Laura J. Meyerson

Hoechst Marion Roussel

Abstract

Randomization, double blinding, and placebo-control methods are all motivated, in part, to control bias. These methods have become the design standards for development of pharmaceutical products. The ability to implement the double-blind procedure with a placebo control in the study of a drug with a distinct safety profile has been questioned in the development of many psychotropic drugs (16). In fact, there are many studies that show through patient and investigator guesses that the blind was not well maintained (4, 7, 9, 14, 20). This paper will seriously question the validity of the double-blind procedure used to study any drug which has a distinct safety profile. Methods for reducing the bias that is induced by the unblinding that naturally occurs will be explored.

The placebo effect is a well-known phenomenon. Placebo is a Latin word meaning "I will please." The placebo effect in the broadest sense usually denotes any improvement in therapeutic measures for agents that lack any pharmacodynamic action on the disease concerned (2, 11). The results are thought to be due to the patient's desire to be healed. It is largely psychologically based and varies between individuals and diseases. Because of the psychological basis, it may be more apparent in subjective measures than objective measures, e.g., questionnaires versus blood chemistry. It is probably the psychological nature of the effect that lends it to be questioned most often by pharmacologists studying psychotropic drugs (16).

Placebo controls are used not only to control for these psychological effects that arise just from being treated, but also, to control for spontaneous symptom change or remission, unrelated to the specific treatment being offered, occurring while in treatment (8). It is thought that for the placebo to control bias adequately, randomization and double-blinding are necessary. The double-blind method denotes masking both the patient and investigator to the actual drug the patient is receiving (whether it is placebo or active). This is done in order to allow the placebo to take its full effect by eliminating possible bias that may occur if the investigator or patient knew that they were receiving an inert agent. If the patient knows to which treatment he has been assigned, he may act consciously or unconsciously in ways that would bias outcome. Likewise, if the investigator is aware of the patient's treatment assignment, she may interact with the patient in a prejudicial manner, or allow his/her subjective opinions of treatment efficacy to influence data collection and evaluation of events occurring during the course of the study (1). Randomization enables the masking of the investigator by not allowing the investigator to determine what treatment each patient receives.

The question is, "Is it really possible to blind both patients and investigators to an effective treatment in a placebo-controlled trial?" This is referred to as Philip's paradox (6). Philip's paradox states that the more potent a therapeutic

variable, the less likely its efficacy can be proven in a double-blind study (6). For example, for a very potent drug with a distinct effect, all patients and investigators would be readily unblinded and thus any result that was seen could be attributed to bias from understanding which compound was assigned and not to the treatment itself. It is important that the investigator and patient know that there is a certain chance (often as high as 50%) that a non-active medication may be administered for treatment in a placebo-controlled clinical trial. In clinical trial research, the topic of unblinding is rarely mentioned (12). This may leave the impression that the double-blind was diligently maintained. This, however, is contradictory to clinical care practices (1). There is natural curiosity taking effect which most studies totally ignore.

The bias incurred from guessing correctly due to potent side effects is different from that due to potent efficacy. In fact, if correct guessing is due to potent efficacy alone, there would be no concern about the results of the efficacy demonstration (7). This is because it is the efficacy itself which is motivating the guess and the response. In contrast, if it is safety (side effects) which motivate the guess to "on active," then the response may be affected in a non-valid way. For example, a patient may think, "Due to the side effects I am experiencing, I must be on active and thus respond in accordance with what would be considered getting better." Therefore, it is important to inquire about what motivates a patient's guess, safety or efficacy.

Hughes and Krahn (5) believe double-blind studies should routinely assess blindness. They reasoned that the blindness of a study is maintained if the frequency of incorrect guesses is greater than the frequency of correct guesses. Secondly, they suggest comparing the magnitude of the drug effect among patients who correctly guessed, incorrectly guessed, and could not tell. If guessing influenced the study's results, then the drug effect would differ for these three groups. They applied this approach to a study of nicotine gum and found there was some unblinding but it did not effect the validity of the study since the drug effect did not differ significantly among the three groups.

Since this factor, a patient's guess, is there and may affect response, like any other covariate, it should be collected and tested for its effect on the primary outcome of the clinical trial. In a survey of clinical trials (6), less than 5% collected these validated data. It would be useful to develop a standard questionnaire for this use. The questionnaire should have some measure of certainty of their answers and reason (side effects, efficacy) for their guess. There should also be a "Don't know" option.

Moscucci, et al. (9) used the following questionnaire to address this issue in a clinical trial of phenylpropanolamine versus. placebo in mild obesity.

1. Do you think you have been on placebo or active medication?
2. How sure are you of your answer (on a visual analogue scale from 0 to 100 with 0 = not sure at all and 100 = completely sure)?
3. Is your impression based on:
   a. Weight loss (yes, no)
   b. Lack of weight loss (yes, no)
   c. Adverse drug reactions (yes, no)
   d. Lack of adverse drug reactions (yes, no)
   e. Appetite control (yes, no)
   f. Lack of appetite control (yes, no)
   g. Other

This questionnaire addresses the motivation issue and also the certainty. It may be confusing to have to choose placebo or active, and it may be better to have a "Don't know" option. It is interesting to note that all patients did choose in this study. This is the type of questionnaire that could be used and the answers would be evaluated as a covariate.

Kirsch and Weixel (10) evaluated the expectancy effects by telling patients they were receiving caffeine when they actually received decaffeinated coffee. They found both subjective responses and physiological responses such as pulse rate performed in accordance with perceived effects of caffeine. They recommend adjusting for these expectancy effects by using a balanced placebo design (17).

	ACTUALLY RECEIVED	
TOLD THEY RECEIVED	PLACEBO	ACTIVE
PLACEBO		
ACTIVE		TRUE MARKET

This design allows independent evaluation of drug effects, expectancy effects, and their interactions, as well as an estimate of the effect as it will be marketed. Practical considerations such as informed consent and sample size sometimes preclude its use.

White, et al. (15) suggest an independent retrospective review of the case report forms, where the reviewer guesses treatment assignment after reviewing therapeutic measures only (not side effect information). Another approach would be to have separate investigators for assessment of therapeutic versus side effects. This would eliminate some of the investigator bias due to knowledge of side effects and therefore drug assignment. The investigator's bias may be greater than the patient's bias, and thus more important to control (19).

This problem with blinding is fairly common knowledge in the development of psychotropic drugs such as benzodiazepines, lithium, and antidepressants (14, 20). Yet, many other products have the potential for bias in their clinical trials. For example, a vaccination for allergy that has a side effect at the time of the vaccine and the efficacy is supposed to be shown 6 weeks later, has this potential for bias. Since the efficacy is also typically shown through a subjective questionnaire rating allergic symptoms, the potential for bias is increased. If the side effects at vaccination have informed the patient which product was administered, the patient and/or investigator are unblinded prior to measurement of clinical response. How does this affect the response? In a long-term trial, it is useful to ask for each patient's guess at various points during the trial. For example, in order to determine association with short-term side effects versus long-term efficacy, one needs to ask at peak drug levels and then at the end of the trial. Other examples where this potential bias has been explored include (-blocker therapy (4), multiple sclerosis therapy (18), appetite depressants (9), and the Aspirin Myocardial Infarction Study (AMIS) (3).

References

1. Zifferblatt, S. M. and Wilbur, C. S. (1978), "A Psychological Perspective for Double-Blind Trials," *Clin. pharmaco. ther,* 23: 1-10.
2. Svedmyr, N. (1979), "The Placebo Effect," *Scand. J. Rehab. Med.,* 11:169-172.
3. Howard, J., Whittemore, A. S., Hoover, J. J., Panos, and the Aspirin Myocardial Infarction Study Research Group (1982), "How Blind was the Patient Blind in AMIS?" *Clin. Pharmacol. Ther.* 32:543-553.
4. Byington, R. P., Curb, J. D., and Mattson, M. E. (1985), "Assessment of Double-blindness at the Conclusion of the beta-Blocker Heart Attack Trial," *JAMA,* 253(12):1733-6.
5. Hughes J. R. and Krahn, D. (1985), "Blindness and the Validity of the Double-blind Procedure," *Journal of Clinical Psychopharmacology.* 5(3): 138-42.

6. Ney, P. G., Collins, C., Spensor, C. (1986), "Double blind: Double Talk or are There Ways to do Better Research?" *Med. Hypotheses*, 21: 119-126.
7. Rabkin, J. G., Markowitz, J. S., Stewart, J., McGrath P., Harrison, W., Quitkin, F. M. and D. F. Klein (1986), "How Blind is Blind? Assessment of Patient and Doctor Medication Guesses in a Placebo-controlled Trial of Imipramine and Phenelzine," *Pshychiatry Research*, 19(1):75-86.
8. Rickels, K. (1986), "Use of Placebo in Clinical Trials II," *Psychopharmacology Bulletin*, 22: 19-24.
9. Moscucci, M., Byrne, L., Weintraub, M. and Cox, C. (1987), "Blinding, Unblinding, and the Placebo Effect: An Analysis of Patients' Guesses of Treatment Assignment in a Double-blind Clinical Trial," *Clinical Pharmacology & Therapeutics*, 41(3):259-65.
10. Kirsch, I. and Weixel, L. J. (1988), "Double-Blind Versus Deceptive Administration of a Placebo," *Behavioral Neuroscience*, 102(2):319-323.
11. Bell, D. S. (1989), "Topics in Clinical Research III. The Importance of Randomized Double-blind Procedures in Clinical Trials," *Clinical Therapeutics*, 11(5):565-7.
12. Oxtoby, A., Jones, A. and Robinson, M. (1989), "Is Your 'Double-blind' Design Truly Double-blind?" (see comments), *British Journal of Psychiatry*, 155:700-1.
13. Munjack, D. J., Brown, R. A., McDowell, D., and Palmer, R. (1989), "Actual Medication Versus Therapist Guesses: In a Blind Study, How Blind is Blind?" (letter), *Journal of Clinical Psychopharmacology*, 9(2):148-9.
14. Margraf, J., Ehlers, A., Roth, W., Clark, D., Sheikh, J., Agras, W. S., and Taylor, C. B. (1991), "How 'Blind' are Double-Blind Studies?" *J. of Consulting and Clinical Psychology*, 59(1): 184-187.
15. White, K., Kando, J., Park, T., Waternaux, and Brown, W. (1992), "Side Effects and the 'Blindability' of Clinical Drug Trials," *Am. J. Psychiatry*, 149:12.
16. Fisher, S. and Greenberg, R. P. (1993), "How Sound is the Double-blind Design for Evaluating Psychotropic Drugs?" (Review), *J. of Nervous & Mental Disease*, 181(6):345-50.
17. Kirsch, I., Rosadino, M. J. (1993), "Do Double-blind Studies with Informed Consent Yield Externally Valid Results? An Empirical Test," *Psychopharmacology*, 110(4): 437-42.
18. Noseworthy, J. H., Ebers, G. C., Vandervoort, M. K., Farquhar, R. E., Yetisir, E. and R. Roberts (1994), "The Impact of Blinding on the Results of a Randomized, Placebo-controlled Multiple Sclerosis Clinical Trial," *Neurology*, 44(1): 16-20.
19. Double, D. B. (1995), "Unblinding in Trials of the Withdrawal of Anticholinergic Agents in Patients Maintained on Neuroleptics," *Journal of Nervous & Mental Disease*, 183(9): 599-602.
20. Morin, C. M., Colecchi, C., Brink, D. Astruc, M. Mercer, J, and Remsberg, S. (1995), "How 'Blind' are Double-blind Placebo-controlled Trials of Benzodiazepine Hypnotics?" *Sleep*, 18(4):240-5.

Standardized Data Structures and Visualization Tools: A Way to Accelerate the Regulatory Review of the Integrated Summary of Safety of New Drug Applications

Jonathan G. Levine and Ana Szarfman
Office of Epidemiology and Biostatistics, CDER, FDA

Introduction

If a safety review by the Food and Drug Administration is to be completed rapidly, it is important that a reviewer be able to efficiently identify in the safety database unanticipated, but serious, adverse events such as agranulocytosis, aplastic anemia, uremic hemolytic syndrome, etc., as well as important drug interactions, and patient subpopulations that are at increased risk of developing adverse events. NDAs often contain a vast amount of data. Without a comprehensive strategy for reviewing the data, a serious side effect may go undetected until the drug is marketed. If the serious side effects are discovered belatedly during the pre-marketing review process, the drug's final approval may be delayed while further analyses or studies are performed.

Is there a way to ensure that all indications of safety problems are identified as early as possible, and thoroughly evaluated as early in the review process as possible? Ideally, information could be assessed as the data accumulates, so that critical issues are found early and timely adjustments could be made, such as lowering the dose, changing laboratory monitoring, etc. in subsequent studies. In practice, there may be a delay in analyzing the data, and studies are done in parallel, not sequentially, so that the modification of studies for safety reasons is not possible without delaying the drug development program. Similarly, small between-group differences in adverse event rates may not become apparent until the results of several large trials are combined.

Safety review questions are hard to specify in advance

Unlike the review of clinical data developed to support efficacy, where the definition of efficacy is prespecified and the study design is usually focused on this hypothesis, safety reviews need to center on the question of whether or not the drug affects some patients adversely, and whether or not the effect is important enough to alter the way the drug is used and labeled. In practice, the process of identifying events, trying to determine differential frequency between treatment groups, and attempting to ascribe causality is loosely defined.

Common sense seems to assure us that it is not a difficult task to find serious adverse events that are fatal, life threatening, disabling, or require withdrawal from the study. However, important rare but serious toxic drug reactions can go undetected, because clinicians may not be prepared to recognize signal cases (Peck, C., Temple, R., and Collins J. M., *JAMA* 269:1550-1552, 1993), the clinical and pharmacologic factors responsible for these events may not be understood (Woosley, R. L. et al, *JAMA*, 269:1532-1536, 1993), risk factors responsible for individual susceptibility may be unknown (Honig P. K., et al, *JAMA* 269:1513-1518, 1993), and a variety of confounding factors may obscure the identification of toxic drug reactions. For example, auto accidents might be dismissed out of hand as a drug related

adverse reaction, when in fact they can be the result of treatment related seizures.

While it also might seem that determining the relative frequency of adverse events would be straightforward, it usually is not. Lack of concurrent controls in open-label extensions of studies, differences in total time patients are exposed to the different treatments, and incomplete dose information are just a few of the problems. Further complicating matters is that, in general, toxic drug events occur as a result of complex, multiple, interdependent biologic and drug induced factors. Although each of the factors per se, including concomitant medical conditions and medications, may not affect the normal drug disposition in most patients, the right combination of factors may trigger unexpected drug reactions. Identifying the important combination of factors that results in the adverse event is no small challenge.

The understanding of the pharmacologic activity of the drugs can help associate the adverse event with a potential cause (Veining, G. R., *British Medical Journal* 286:289-292, 1993; Szarfman, A. et at, *New Engl. J. Med.* 332:193, 1995) and may also help formulate better safety hypotheses a priori. This is especially true when the safety profile of other drugs in the same class is understood, or additional data on the drugs is available. Currently, access to the standardized post-marketing Safety Reporting System (SRS) database is giving us such an opportunity. The SRS adverse event database is a large (over 1.2 million primary records) relational database containing all of the structured data from spontaneous adverse event reports received by FDA during the postmarketing period.

In this article, we discuss some of our own experiences as clinical and statistical reviewers at the FDA, and our experimentation with new tools to understand safety. These approaches have been tested by Ana Szarfman in the review of safety of an NDA and in the review of the post-marketing Safety Reporting System (SRS) database.

Evaluation of a safety database requires hands-on work by clinicians and statisticians

Studies that are statistically powered to detect differences in primary efficacy endpoints are usually not designed to detect differences in safety outcomes. At the end of a sponsor's drug development program, we are invariably left with data not adequately powered to test hypotheses about rare safety events. While standard statistical tests and estimates are often included in submissions, classical statistical methods cannot test for specificity and causality of unanticipated adverse events because the sample sizes are too small for conventional analyses. Because of this, a safety review is inevitably a combination of art and science.

A good safety review requires hands-on work by both clinicians and statisticians. Since clinicians, typically, do not know how to do the programming needed to perform the kinds of complex analysis needed to investigate potential safety issues, they rely on statisticians and programmers to perform these analyses. Unfortunately, the work of the person/group doing the programming is complicated by the fact that safety review questions are hard to specify a priori. What often happens is that the analyst produces volumes of static reports, resulting in reviewers poring over volumes of line listings, summary statistics, and generic graphics. These static reports force the reviewer to follow somebody else's analysis, are often difficult to interpret, and almost always fail to address several key medical questions.

An alternative strategy is for the reviewer to do their own analysis from sponsor-provided data sets. More often than not, data sets need to be modified in order to perform desired analyses. This requires extensive new programming that cannot readily be reused. Moreover, reviewers' analyses must cope with many hurdles and navigate many mazes: data cleaning, idiosyncratic data representations, transformation schemes, conversion of diagnostic methodologies and measurement units, combining heterogeneous data, and the all important need to keep track of all these steps. Although statisticians have the skills to carry out these steps, it is not an optimal use of a statistical reviewer's time, and is prone to error. Medical reviewers, while trained at recognizing potential toxic drug reactions, are generally not able to perform these tasks and may not appreciate the potential noise that these types of data manipulation can add to the statistical summaries. Communication between the groups assigned to analyze the safety data is an additional obstacle and all these factors slow the review process. These factors unnecessarily complicate the analysis of safety signals requiring hands-on work by clinicians and statisticians, and hinders building upon previous analyses.

Safety data presents unusual challenges

Traditional statistical approaches for efficacy analysis emphasize hypothesis tests and estimates of measures of central tendency from designed studies. In safety review, measures of central tendency are often the least important aspect of the data. The average response often has little predictive value. Instead, identifying a small number of extreme values in patients can be the most important task for the safety reviewer.

Combining safety data across studies can help detect the more serious, rare adverse events and extreme values that occur at a low frequency. Combining safety data from different studies is especially important because most individual clinical trials have inadequate power to evaluate safety. Combining these data is not straightforward, since the data is often derived from multiple independent groups within a pharmaceutical company, multiple contract research organizations, or academic institutions who design studies and collect information using their own stand-alone systems, all of which complicates the process of combining studies when no standard approach exists.

To date, standards for the collection, representation, and organization of safety information submitted have not emerged. This results in the submission of NDAs that are set up to do study-by-study review. These safety data files are difficult to merge into a single database. While this may not be a major problem with efficacy databases (since the statistical approach is to analyze separate studies), this lack of a consistent data structure across studies severely hinders the efficient review of safety. Typically, a reviewer has to cope with a succession of study-specific data representation and transformation schemes (methods conversion, units conversion, etc.) before the data can be combined and re-analyzed. It is not unusual for as much as 80-90% of the review time being spent modifying the data in order to do the needed analysis and only 10-20% of the time actually doing the analysis.

Once the safety data is combined across studies, the data will invariably need to be subsetted. Adequate subsetting of the data can help identify potentially susceptible patient subpopulations, including patients with renal and/or liver function abnormalities, the very old and the very young, women or men, low weight patients, patients taking concomitant medications, etc. Susceptible subgroups can be detected by discovering systematic increases of specific adverse events that are dose dependent, or dependent on a specific concomitant condition or medication. For example, the presence of systematic dose dependent phototoxic reactions in patients working outdoors vs. its absence

in hospitalized patients, not exposed to the sun, should trigger an alert. Similar analysis should be extended to withdrawals due to adverse events, serious adverse events including deaths, and clinical laboratory adverse events. Subsetting by drug concentration, when available, can help reveal further safety problems.

In any safety analysis, outliers need to be highlighted. It is essential that extreme values not be excluded from an analysis except when they are due to data errors, since skilled clinicians may find them to be the first alert to important toxic drug effects. For events that are very rare in the general population, the discovery of these events in patients exposed to the test drug during drug development, often is taken as an evidence of causality until the diagnosis of drug toxicity can be confidently excluded.

In all cases, causes of outliers need to be understood, since failure to adequately characterize the effect of the drug in outliers can lead to a delay in understanding the safety profile of a drug and its risk/benefit ratios. Understanding the causes of outliers can help the drug development program adapt to new study designs; define patients characteristics, patient outcome, lab values, etc.; and make early and timely adjustments, such as lowering the dose, reducing the duration of treatment, changing clinical laboratory monitoring, etc. in subsequent studies. Because

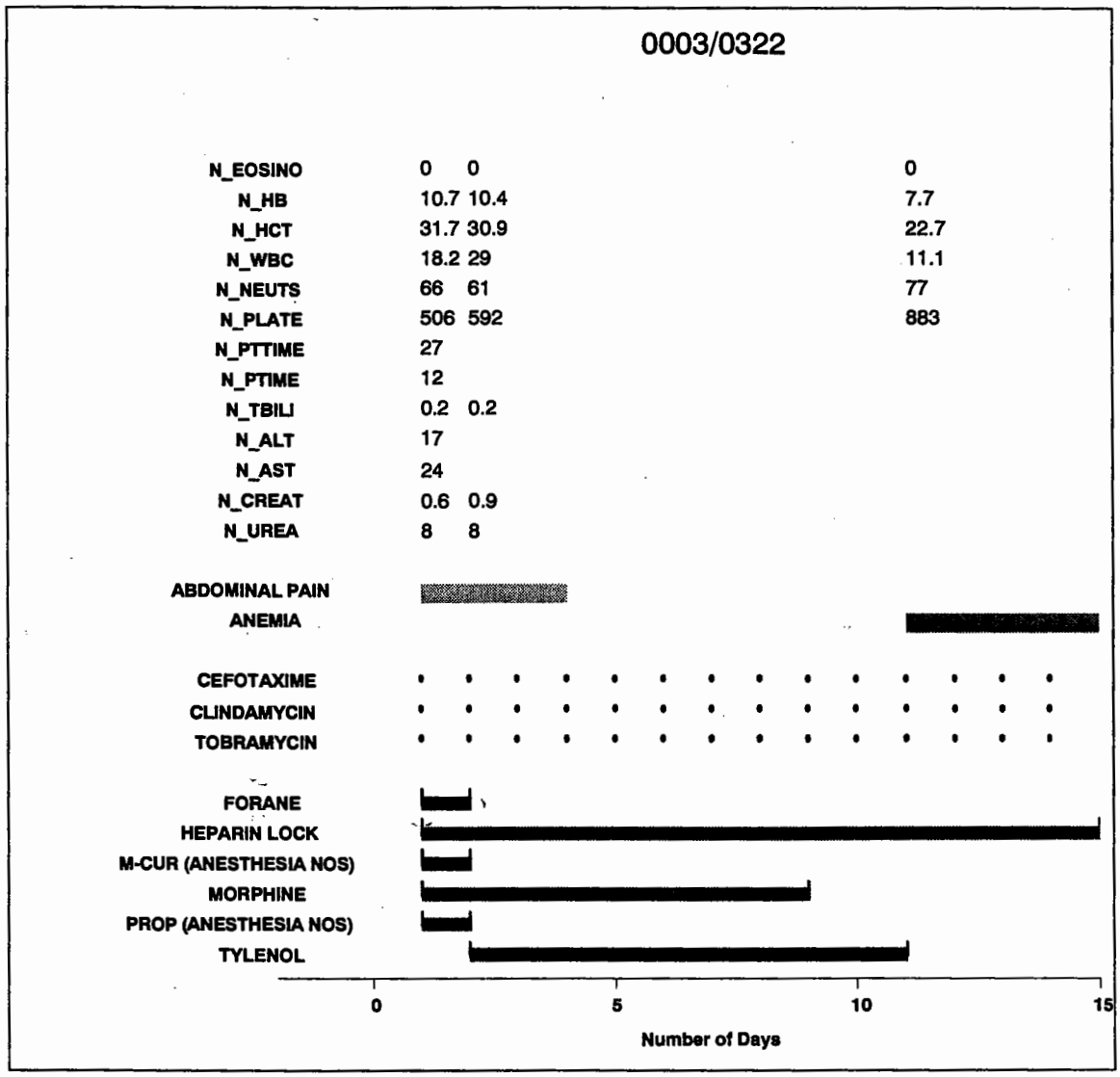


Figure 1. Patient Timeline Summary

of this, "drill down" capability (the ability to access additional data about the patient with the extreme value) is invaluable.

Another challenge in safety review is that a small number of data errors can obscure important safety results. If two patients are miscoded as having neutrophils of 4000 (103/uL) instead of 400, a potentially life-threatening adverse event may be overlooked until the drug is marketed. It is worth noting the opposite case, where patients with neutrophils of 4000 are miscoded as 400 and it is more likely to be detected and corrected. It is surprising how often errors of both types are found by FDA reviewers. These errors disrupt the analysis process unnecessarily.

Visualization tools are a natural way to explore safety data

The utility of graphical techniques in understanding complex data is widely recognized. While no one would consider presenting a map of the world exclusively in tabular form, it is standard to present safety results as a series of tables. We have found that visual displays are invaluable in understanding safety data. The additional organization is critical for the identification of real patterns of clusters and outliers, (including the identification of unanticipated adverse events), trends, and correlations.

Graphic tools can also aid the data validation process by

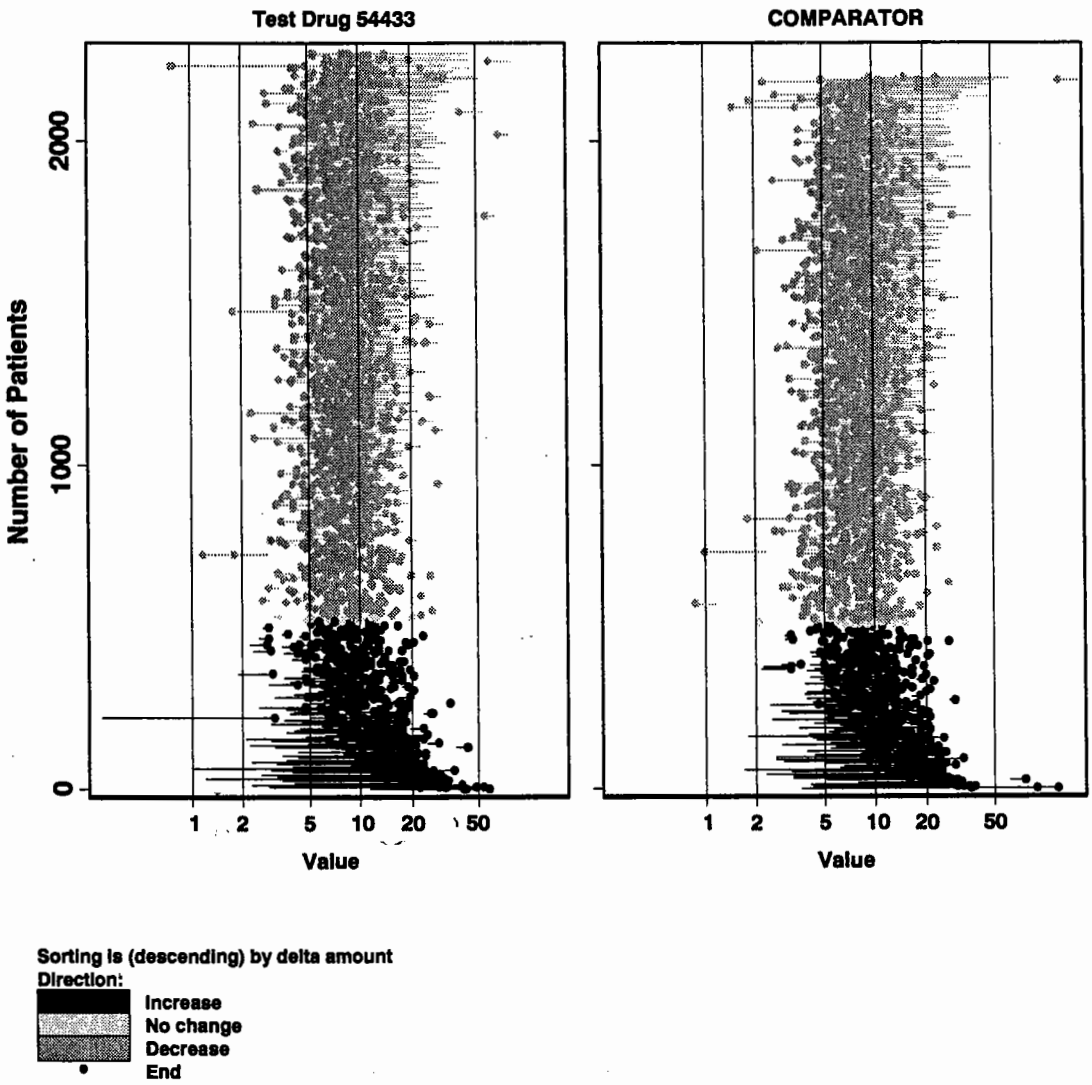


Figure 2. Delta Plot

identifying the abnormal values that have to be validated. Graphics can also help identify errors of data collection and merging. For example, the erroneous merge of two different units of measurements (prothrombin time in seconds and as percent of the control) will appear as bimodal distributions or as clusters of outliers; these features are lost in tabular displays. The problem that we all have to face is that, similarly, safety signals can be lost within standard output tables.

There are multiple advantages of using standardized data structures of the whole NDA data if we wish to use visualization tools for exploratory and confirmatory purposes. It is much faster to compare study results after merging the data from all studies and partition the data, than to look at each study and each variable in isolation and then try to understand the differences between the different variables.

Once the data from all studies is integrated it is very easy to visually explore the data from several perspectives by partitioning the visual displays by trial, gender, weight, serum creatinine at baseline, indication, treatment, dosage, formulations, treatment duration, outcomes, concomitant medications and conditions, etc. This flexibility speeds the reanalysis and identification of subpopulations that might be at a higher risk or the discovery of differential safety of the drug. It also improves the quality of the statistical analysis because it is not applied blindly. Data can be

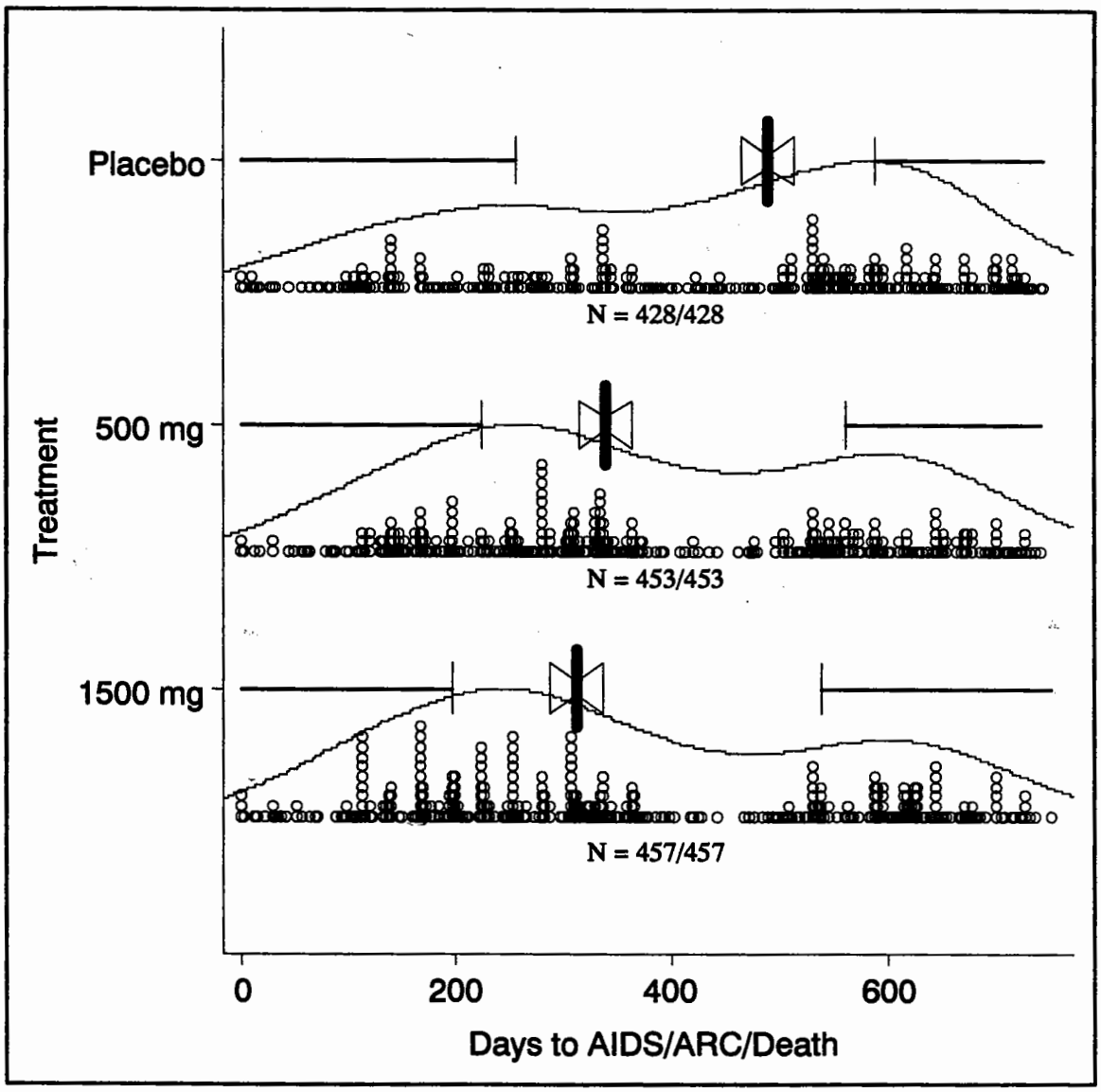


Figure 3. Box plots with average shifted histograms

subsetted in a variety of ways, including coding groups by color, symbol type etc., and the use of multiple aligned displays.

Graphics, when included in an NDA, are static plots of a subset of all possible displays. The choice of displays, no matter how well intentioned or planned, may lead to real or perceived obfuscation of the data. An alternative is to use visualization tools that are interactive. This requires a computerized visualization system configured to display clinical trial safety data and the electronic submission of associated data in an integrated structure.

Clinical safety data are multivariate and many factors are interdependent. No single display of information can take this into account. Static displays of one variable or two variables at a time can give only a very partial display of these complex relationships. Understanding can be further increased by the addition of interactive tools that enable interactive exploration and comparison of the whole NDA data. By being able to condition on several stratifying factors like trial, dose, weight, gender, or concomitant conditions, the understanding of the multiple facets of the data is made more transparent.

The Center for Drug Evaluation and Research (CDER) of the FDA is pilot testing new interactive computer graphics-based systems for summarizing and graphically displaying the integrated laboratory database for the entire NDA patient and subject population under the Computer-Assisted Review of Safety (CARS) project.

The goal of CARS is to facilitate the access to the laboratory and safety information. The CARS group is working to define a database format standard. Standardized data structures need to be implemented so that data from multiple sources can be analyzed with little or no programming by the reviewer because the system will take care of that. This will facilitate the integration of databases across studies and the systematic application of tools that can produce standardized interactive graphic displays of the data and will enable reviewers to visualize and recognize the presence or absence of patterns of abnormalities without the need of complicated programming.

Examples

Figure 1 shows an example of a "Patient Timeline Summary Graph" used in the review of an NDA. In this graph we can display and link clinical laboratory measurements, adverse events, and medications on a common timeline, one graph per patient, for the patients who died or dropped out of the study for AE's.

Figure 2 displays a "Delta Plot" for patients' baseline and posttreatment values. In this graph, each line represents an individual patient, and the ends of each line represent the pre- and post-treatment measurements. This type of display can convey much of the information provided by a scatterplot while avoiding the problem of overplotting. Also, by judiciously choosing the sorting variables used in creating the display, we can identify and drill down to important cases.

In Figure 3 we are simultaneously presenting the raw data, a nonparametric density estimate of the data (an average shifted histogram; see Scott D. W., *Multivariate Density Estimation*, John Wiley, 1992) and a boxplot. At the bottom of each boxplot is the number of nonmissing data values and nonmissing data values + missing data values.

High Quality, Standardized, Data Structures: The Key to Interactive Visualization Tools

If interactive visualization tools are to be used in a review, it is imperative that data used in the display be readily available in a computerized electronic data structure. Equally important is the need for data structures that can be combined without requiring additional programming efforts. Perhaps less well recognized is the need for the inclusion of "meta data" in the data structures. Information concerning definition of study visits and dose specified in the protocol (as opposed to dose actually received) is often unavailable in the original computer data submitted. If it is available, it may be inaccessible. In practice, the hardest part of producing and evaluating a visual display is often formatting the data.

We cannot overemphasize the importance of high quality data. Efforts at improving the quality of the collection and recording methods are worthwhile, because a handful of data errors can disrupt the review process, especially when important safety signals might be present in a frequency that is lower than the collection noise.

Conclusions

In the past five years, the advances in computer technology, data access, and interactive data visualization have been dramatic. These advances make the implementation of interactive graphic tools both possible and cost efficient. Statisticians and clinicians will benefit from applying them during the drug development and review processes.

The lack of integration of the safety database works against efficiently understanding potential safety issues because the evaluation of the data as a whole entity is hindered. Data are frequently analyzed too late for the results to have an impact on new studies and this affects the drug development and submission time, and the quality of the drug development program (Szarfman, A., *Regulatory Affairs Focus* 1:12, 1996).

Interactive visualization tools, in concert with an integrated database, speeds data analysis and facilitates the understanding of the safety data. In order to produce quality visualization outputs, we need to invest in improving the collection of "cleaner" data and in building high quality data structures that can be used to interrogate the data without cumbersome and hard-to-document manipulations. This will help focus the effort away from costly data manipulations and toward the timely understanding of the data.

Acknowledgment:

In this article, we discuss some of our experiences as FDA clinical and statistical reviewers and our experimentation with new tools to understand safety data. This article expresses only our own views and does not reflect CDER's policy in this area. We acknowledge the increased insight we have gained not only as hands-on reviewers but also from our participation as members of several Good Review Practices groups initiatives and from discussions with members of these committees and with Robert O'Neill.

Workshop on the Collection and Analyses of Adverse Events Data

The workshop is designed to consider problems in the definition, collection, reporting and analyses of adverse events data during the development of a new pharmaceutical product. Issues such as the collection of symptoms vs. syndromes, time window of event collection after treatment discontinuation, collection of adverse events at baseline and in long-term trials, analyses of laboratory data, U.S. and European regulatory perspectives on the collection and summary of the adverse events data, and alternatives to current approaches will be discussed.

Although the workshop is sponsored by the Biopharmaceutical Section of The American Statistical Association, we encourage attendance by our non-statistician colleagues in the medical and regulatory areas.

Location: This Workshop will be held at the Hyatt Bethesda, Bethesda, Maryland. This hotel is located on the Metrorail line just six miles from the U.S. Capitol and convenient to over 100 restaurants within walking distance.

Hotel Reservations: Hotel rooms are available at the single or double rate of $125.00 plus tax. There is a limited number of government rate rooms available at the prevailing government per diem of $112.72. To make reservations call 301-657-1234. These rates are available until October 2, 1996.

Preliminary Program

Monday, October 28

7:15 a.m.-8:00 a.m. **Continental Breakfast**

8:00 a.m.-8:30 a.m. **Overview of Issues Concerning Adverse Events Data**

— Speaker: ROBERT STARBUCK, Wyeth-Ayerst Research

8:30 a.m.-9:30 a.m. **Adverse Events: Definition, Collection and Standard Summarization**

- Chair: ROBERT NORTHINGTON, Wyeth-Ayerst Research

— Speaker: ROBERT NORTHINGTON, Wyeth-Ayerst Research
— Speaker: THOMAS COOK, Merck Research Laboratories

9:30 a.m.-10:00 a.m. **Morning Break**

10:00 a.m.-12:00 noon **Laboratory Data - The Usual and the Special Analyses**

- Chair: THOMAS LIN, Sandoz Research Institute

— Speaker: ROCCO BRUNELLE, Eli Lilly & Company
— Speaker: STEVE NETTLER, Sandoz Research Institute
— Speaker: D. CRAIG TROST, Pfizer Inc.

12:00 noon-1:30 p.m. **Lunch** (on your own)

1:30 p.m.-2:45 p.m. **Why Aren't AE Collection/Recording/ Summarization Straightforward?**

- Chair: CURTIS WILTSE, Eli Lilly & Company

— Speaker: CURTIS WILTSE, Eli Lilly & Company
— Speaker: LAURA MEYERSON, Hoechst Marion Roussel

2:45 p.m.-3:15 p.m. **Afternoon Break**

3:15 p.m.-5:00 p.m. **Special Topics in Adverse Event Analysis and Reporting (e.g., Correlating AE with PK Data, Special Challenges for Long-Term Trial, Benchmarking of AE Procedures Among the Industry)**

- Chair: SALLY GREENBERG, Berlex Laboratories

— Speaker: ROBERTA SMITHEY, Eli Lilly & Company
— Speaker: LOTHAR TREMMEL, Amgen, Inc.
— Speaker: MICHAEL HALE, Roche Global Development

5:00 p.m.-7:00 p.m. **Reception**

Tuesday, October 29

7:15 a.m.-8:00 a.m. **Continental breakfast**

8:00 a.m.-10:00 a.m. **Regulatory Perspective on the Analyses and Presentation of Adverse Events Data**

- Chair: TONY SEGRETI, Glaxo-Wellcome, Inc.

— Speaker: ANA SZARFMAN, Food & Drug Administration
— Speaker: To Be Announced, Food & Drug Administration
— Speaker: STEPHEN EVANS, Medicine Control Agency

10:00 a.m.-10:30 a.m. **Morning Break**

10:30 a.m.-12:30 p.m. **Can We Find a Better/Alternative Solution (in Areas of ISS, Risk vs. Benefit in Safety Assessment, the Use of Hazard Function in Safety Assessment)?**

- Chair: ED LAKATOS, G. D. Searle

— Speaker: NANCY SILLIMAN, Food & Drug Administration
— Speaker: JAY HERSON, Applied Logic Associates, Inc.
— Speaker: DAVID SALSBURG, Salsburg Statistical Consulting

12:30 p.m.-2:00 p.m. **Lunch** (on your own)

2:00 p.m.-3:30 p.m. **Panel Discussion**

- Moderator: CHRISTY CHUANG-STEIN, Pharmacia & Upjohn, Inc.

- Panelists:
  - ROBERT NORTHINGTON, Wyeth-Ayerst Research
  - D. CRAIG TROST, Pfizer Inc.
  - CURTIS WILTSE, Eli Lilly & Company
  - ANA SZARFMAN, Food & Drug Administration
  - STEPHEN EVANS, Medicine Control Agency
  - DAVID SALSBURG, Salsburg Statistical Consulting

3:30 p.m.-4:00 p.m. **Afternoon Break and Workshop Adjourns**

Section News

Deadline for Proposing Invited Paper Sessions

Tom Capizzi
Merck

The Biopharmaceutical Section sponsors invited paper sessions at the annual International Biometric Society ENAR Spring Meeting and the annual Joint Statistical Meetings. Anyone who wishes to organize an invited paper session for one of these meetings in 1998 should contact the 1998 Program Chair, Tom Capizzi, by June 1, 1997 for the ENAR meeting, or by July 1 for the JSM.

Thomas Capizzi, Ph.D.
Director of Clinical Statistics
CBARDS
RY 33-404
Merck Research Labs
Rahway, NJ 07065-0900
phone: (908) 594-4202
fax: (908) 594-6075
E-mail: tom_capizzi@merck.com

Stan Schor Receives Career Achievement Award

Tony Segreti
Glaxo Wellcome

Stanley S. Schor received this year's Career Achievement Award from the Biostatistics Committee of the Pharmaceutical Research and Manufacturers of America. This award has been given since 1990 to individuals who have made significant contributions to biostatistics and its application in the pharmaceutical industry. Schor was recently honored at a banquet of the Midwest Biopharmaceutical Statistics Workshop.

Schor had a distinguished career at Merck & Co. from 1975 until his retirement in 1990. He served as Executive Director of Clinical Biostatistics and Research Data Systems and led 225 statisticians, epidemiologists, data managers, and computer scientists. Under his guidance, the Merck group was a leader in establishing the independent role of biostatistics in the pharmaceutical industry, initiating the disciplines of pharmacoepidemiology and health economics in the industry, and encouraging staff development through publications, presentations, and service in professional societies.

Prior to joining Merck, Schor served as the Director of Biostatistics at Chicago Medical School and Professor and Chairman of the Department of Biostatistics at Temple University. He was also a long-time faculty member at the University of Pennsylvania. He was elected a Fellow of the ASA in 1972.

Schor is currently enjoying his retirement in Florida where he reports he is playing softball every day.

Biopharmaceutical Section Student Paper Awards

Lianng Yuh
Pfizer Central Research

One of the highlights of the Biopharmaceutical Section Business Meeting on August 6, 1996 was the presentation of student paper awards. The five winners of this competition were:

Li Chen, Department of Biostatistics, Harvard School of Public Health, *Analysis of Multivariate Survival Times with Non-Proportional Hazards Models.*

David Dunson, Department of Biostatistics, Emory University, *Dose Dependent Litter Size and Implications in Quantitative Risk Assessment for Developmental Toxicity.*

Karen Higgins, Department of Biostatistics, Harvard School of Public Health, *The Effect of Serial Dilution Error on Calibration Inference in Immunoassay.*

Qi Zeng, Department of Biostatistics, Harvard School of Public Health, *Bootstrap 'Calibrated' Calibration Confidence Limits for Immunoassay.*

Hongwei Zhao, Department of Biostatistics, Harvard School of Public Health, *A Consistent Estimator for the Distribution of Quality Adjusted Survival Time.*

The Biopharmaceutical Section is pleased to recognize this year's winners, each of whom received an award of $1000 and a plaque recognizing the paper and the author.

The purposes of the student paper awards are to encourage the study of statistics and its practice in the biopharmaceutical industry, and to increase student participation in the Section's programs and activities at the Annual Joint Statistical Meetings.

It is not too early to begin thinking about the 1997 Biopharmaceutical Student Paper competition. In order to be eligible for an award, the student must be:

- an ASA member (or join at the time of abstract submission);
- a degree candidate during the 1995-1996 and/or 1996-1997 academic year(s) at an accredited institution;
- the first author of the abstract;
- willing to attend the 1997 Annual Joint Statistical Meetings to present the paper.

An additional requirement is that the abstract must be submitted to the Biopharmaceutical Section and included in one of its sponsored contributed paper sessions. In addition to meeting the ASA requirements for abstract submission, the nomination procedure also requires the submission of the abstract, the manuscript, and endorsements from the student's advisor and department head by June 1, 1997 to the Section Program Chair:

Lianng Yuh
Director, Department of Biometrics
Pfizer Central Research
Eastern Point Road
Groton, CT 06340
phone: (860) 441-1531
fax: (860) 441-3600

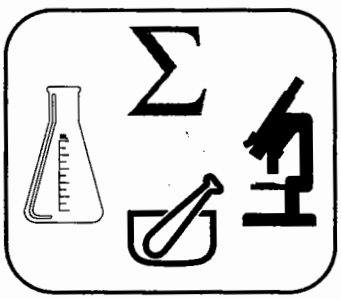
We Need Editors!!

The Biopharmaceutical Report has become a very useful means of getting information out to pharmaceutical statisticians. To maintain the high quality of the Report, we will now have three editors: an editor-elect, editor, and past editor. Each editor will serve three years, with the first year serving as an apprenticeship period. We need volunteers to serve as editors. If you have an interest in this position please contact Bob Davis, Astra Merck Inc. at (610) 695-1070, Fax (610) 695-1961 or E-mail bob.davis@astramerck.com.

Let's Hear from You!

If you have any comments or contributions, contact Co-Editors William J. Huster, Eli Lilly and Company, Lilly Corporate Center, 2233, Indianapolis, IN 46285; phone: (317) 276-9802; fax: (317) 277-3220; E-mail: huster@lilly.com or Curt Wiltse, Lilly Corporate Center, 2233, Indianapolis, IN 46285; phone: (317) 276-5773; fax: (317) 277-3220; E-mail: wiltse_curtis_g@lilly.com

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



Biopharmaceutical Report
c/o American Statistical Association
1429 Duke Street
Alexandria, VA 22314-3415
USA

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

FIRST CLASS POSTAGE