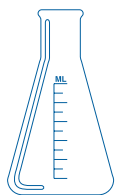


## Biopharmaceutical Section



American Statistical Association

# Biopharmaceutical Report

Volume 8, No. 2

Fall 2000

Chair: *Tom Capizzi*

Editors: *Demissie Alemayehu, Kannan Natarajan, and Ersen Arseven*

## Editor's Note

Demissie Alemayehu

This issue of the Biopharmaceutical Report features a paper on the timely topic of data mining. We hope the paper will provoke further discussion among statisticians in the biopharmaceutical community, and lead to a better understanding of the issues raised by the authors.

An upcoming issue of the Biopharmaceutical Report will feature a paper on challenges and opportunities of Phase IV statistics. The topic will be a natural extension of the themes of earlier issues (cf. Vol. 7, No. 2 and Vol. 7, No. 3). We have been encouraged by the positive feedback we have received on the topics covered in previous issues. However, if you have ideas for a topic that would be of interest to our readers, you are welcome to contact one of the editors.

Finally, please make sure to visit the Biopharmaceutical Section's Web site to get current information on the Section's activities, including workshops, awards, membership listing, and past issues of the Biopharmaceutical Report. The site can be accessed at [www.best.com/~asabp](http://www.best.com/~asabp).

## Knowledge Discovery from Databases and Data Mining: New Paradigms for Statistics and Data Analysis?

Herman P. Friedman, M. Phil. and  
Judith D. Goldberg, Sc.D.

*Herman P. Friedman is President, Statistical Science and Technology Associates, Inc.; Judith D. Goldberg is Professor and Director, Biostatistics Division, New York University School of Medicine, 650 First Avenue, 5th Floor, New York, New York 10016 (address for reprints).*

### Abstract

In this paper, we examine the question of whether or not data mining and knowledge discovery from data is intrinsically different from statistics and data analysis. We provide a description of the origins of data mining and knowledge discovery from data from the data mining perspective and

## Contents

### Editor's Note

..... ALEMAYEHU 1

### FEATURED ARTICLE

#### Knowledge Discovery from Databases and Data Mining: New Paradigms for Statistics and Data Analysis?

FRIEDMAN AND  
..... GOLDBERG 1

### Letter from the Chair

.....CAPIZZI 11

### BIOPHARMACEUTICAL SECTION NEWS

#### Minutes of the August 15, 2000 Executive Committee Meeting

.....CAPIZZI 12

#### 2001 ASA Election Candidates

.....SNAPINN 14

#### Joint Statistical Meetings Best Contributed Paper Awards

.....ROE 14

#### The 24th Annual Midwest Biopharmaceutical Statistics Workshop

.....NEZAMIS 15

#### 2000 ASA Fellows

.....GOULD 15

from statistical perspectives. The three classes of methods that dominate the data mining literature are reviewed. Two of these methods are classification methods and the third are methods and procedures for searching for interesting relationships. Many of the algorithms associated with these approaches come from the traditional disciplines of statistics, machine learning, neural networks, and artificial intelligence; methods for evaluation are primarily from statistics. We also examine statistics and data analysis from a global point of view that includes formal and informal principles for relating data, theory, and context. We conclude that data mining and KDD can be accommodated within the broad view of data analysis and statistics. Many opportunities have been created for the data analyst described by Dempster (1971, see quote on next page). In particular, we identify opportunities for exploratory analysis of nonexperimental data for statisticians in the pharmaceutical industry. Selected references are provided for the reader to explore further the issues that are raised here.

The data analyst envisaged here is a professional who sits in a central position among investigators with data, theoretical statisticians and computer specialists. Among these traditional scientific types, he should serve the objectives of communication and integration. His direct contributions are to the development of new techniques and to the accumulated experience in the use of many techniques. For this, he needs the wisdom to evaluate proposed techniques along dimensions of efficiency and resistance to error, both statistical and computational, and along the dimension of relevance to the substantive scientific enterprise involved.

[A. P. Dempster, 1971, p.317]

## 1.0 Introduction

Today, the ability to collect, store, and retrieve amounts of data that were unthinkable in the past has led to an explosion of interest in the analysis of data that reside in various databases that were collected for many purposes. How to extract “information” and “knowledge” from such databases or “data warehouses” has become big business. And data analysis, as we thought we knew it, has become “data mining”. Statistical software has become data mining software (cf., *Amstat News*, June, 2000). Our tools, S-PLUS®, SAS®, SPSS®, are now being marketed as data mining tools.

In *The New York Times*, we read that “Boston U. Finds a Company to Sift Landmark Heart Data” (June, 17, 2000, p. A9). The new company, Framingham Genomic Medicine, will use the data accumulated since 1948 from the Framingham Study including genetic data collected more recently and “assemble the study’s data in ways useful to

companies or scientists. *Science* (V.288, June 30, 2000, pp 2301-2302) describes a “University Company to Exploit Heart Data”, that is “a bioinformatics company that will mine the data”. A comprehensive database will be constructed followed by the work to “correlate clinical records with DNA analyses from blood samples on file, with the goal of identifying some 50,000 genetic markers in individuals with that are linked to specific abnormalities or diseases.” We note that this landmark follow-up study has been analyzed by statisticians over many years and has yielded important observations with respect to risk factors for coronary disease and new statistical methods (e.g, logistic regression). Now, it is to be ‘mined’!

What happened? Should statisticians be concerned? What, if anything, is lost? The goal of this paper is to help the reader to answer these questions and to provide some selected references and points of view, and to identify some challenges, and opportunities for statisticians in the analysis of non-experimental data in the biopharmaceutical industry.

## 2.0 What is Data Mining?

### 2.1 Origins of “Data Mining” and “Knowledge Discovery from Databases”

The efficiency with which a query can be processed against a database depends upon the organization of the physical database. Items that are physically close can be easily accessed together. The problem that arises is that at the time of the physical database design, the queries that will later be requested are often unknown. Early research at IBM created data streams by monitoring the queries. Cluster analysis was then used to analyze these data streams to reorganize the physical database (Gorenstein, 1974). In fact, while studying for his Ph.D., Piatetsky-Shapiro, studied how a database could optimize itself through machine learning in 1986 (2000).

As organizations grew in size from expansion or mergers and acquisitions, the databases within an organization grew in size, number, and complexity. Database access became very inefficient and the ability to link records from different databases within an organization became very difficult, if not impossible.

The database suppliers then introduced data warehouse products. Data warehouses are essentially databases that can be organized in a modular fashion to meet the needs of the organization. Thus, instead of attempting to reorganize and link ‘legacy’ databases, the data warehouse was created.

The term ‘data mining’ was introduced by the database suppliers to describe software that would search and/or query the data warehouse for interesting and relevant rela-

tionships. Of course, interesting and relevant had to be defined a priori.

Our guess is that the person or persons who coined the term 'data mining' were not cognizant of pre-existing terms with negative connotations such as 'data dredging'.

Piatetsky-Shapiro (2000) organized a workshop in 1989 and was struggling to name it. He rejected the name 'data mining' which was already in use in the database community because it seemed "unsexy" (p.1) to him and was viewed negatively by statisticians. He credits himself with creating the moniker 'Knowledge Discovery in Databases' (KDD) focusing discovery on knowledge. KDD became popular in the Artificial Intelligence and Machine Learning communities; but, "the database researchers were on better speaking terms with the business folks and the press, and the term 'data mining' became much more popular with the business press." (p.1). Piatetsky-Shapiro also predicts that "the data mining industry will overcome the hype stage, and will merge with the database industry" over the next 10 years (2000).

Knowledge discovery from databases has been defined by Fayyad, et al (1996, p.6) as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." In the same paper, data mining is defined as "a step in the KDD process consisting of particular data mining algorithms..." (p.9).

## 2.2 As Viewed by Statisticians

The *Encyclopedia of Statistics*, Update Vol 3 (Lovell, 1999), defines data mining as referring "to the exaggerated claims of significance and or forecasting precision generated by the selective reporting of results obtained when the structure of the model is determined experimentally by repeated applications of such procedures as regression analysis to the same body of data. . . synonymous with 'data scrubbing', data fishing, Darwinian econometrics (survival of the fittest)."

What follows are several descriptions of data mining from the recent statistical literature.

David Cox (1995) notes that "Often the term data mining is used in a rather derogatory sense... I understand mining to be a very carefully planned search ... not a haphazard ramble. Mining, is thus a rewarding but of course dangerous activity. It is an interesting issue, very specific to each subject matter field, as to what extent important conclusions from data 'lie on the surface'."

David Hand (1998) describes data mining "as a new discipline, lying at the interface of statistics, database technology, pattern recognition, and machine learning, and concerned with the secondary analysis of large databases in order to find previously unsuspected relationships which are of interest to the database owners." This is surprising,

since, in his book, *Construction and Assessment of Classification Rules* (1997), he regards "building allocation rules, supervised pattern recognition, discriminant analysis, whatever name you like to use for the activity, as a paradigmatic *statistical problem*." (preface) He also recognizes recent contributions from Computer Science (machine learning) and concludes that classification is an interdisciplinary field. In fact, Hand has a forthcoming book "*Principles of Data Mining*".

Huber (1997, p.304) points out that "On the whole, the data mining community, mostly coming from data base management and logic backgrounds, does not yet seem to be sensitized to the specific problems arising with statistical data, where relationships hold only on the average, and where that average can be distorted by selection bias or similar effects."

DuMouchel (1999) views the search for associations in large databases as one of the tasks of data mining and applied statistical modeling approaches to screen the FDA spontaneous reporting database. We will discuss this further in Section 5.2.

It is interesting to note that statistical considerations were recently discussed at KDD meetings and recommendations made such as the need for new data, for cross-validation, for adjustments for multiplicity, and for use of resampling and randomization methods (Jensen, D. (2000): *Data Snooping, Dredging and Fishing: The Dark Side of Data Mining*). The reader is referred to the SIGKDD website (<http://research.microsoft.com/datamine/sigkdd/>).

## 3.0 Some Methods of Data Analysis Used by Data Miners

Three classes of methods dominate the data mining literature. Two of these methods are classification methods and the third are methods and procedures for searching for interesting relationships. Many of the algorithms associated with these approaches come from the traditional disciplines of statistics, machine learning, neural networks, and artificial intelligence; methods for evaluation are primarily from statistics. Goebel and Gruenwald (1999) provide a fairly exhaustive survey of the software tools used by data miners.

### 3.1 Classification Methods When the Classes are Known

The problem is to determine classification rules based on the data and then to assess the performance of the rules. The usual approach is to determine the rules using a training set of data and to assess performance on a different

dataset. Statistical thinking is useful in the development of an appropriate training set, in screening possible rules for further evaluation, and in the development of statistical procedures and criteria for the assessment of reliability, performance, and generalizability of the rule to new data. The issues are analogous to those in variable selection for regression. In machine learning terminology, the computer is learning from the training set which variables or features are important in distinguishing the known classes (supervised learning).

Neural networks are also being used. However, the performance of the neural network is conditional on knowing the relevant features. In most instances, other methods are used to identify the features, and the classification rule is then developed by the neural network from the training set.

Some useful references that describe classification methods with known classes, including methods of recursive partitioning, are: Ripley (1996), Hand (1997), McLachlan (1992), and Breiman, et al. (1984).

In a recent technical report, J.Friedman, et al (1999) provides an analysis of a procedure from machine learning called 'boosting'. Boosting is a procedure for developing a high performance classification rule from many low performing rules. In words of the authors, "we show that this seemingly mysterious phenomenon can be understood in terms of well known statistical principles, namely additive modeling and maximum likelihood." This provides an excellent example of the interplay of between the development of rules and understanding their behavior.

### 3.2 Classification Methods When the Classes are Unknown (Cluster Analysis)

When the classes are unknown, the classes need to be determined from the data. Cluster analysis is a term used to describe loosely connected concepts, criteria, and algorithms that mostly result in partitions and/or hierarchical groupings of observed data. Algorithms for cluster analysis come from many sources including statistics, machine learning, pattern recognition, biology, engineering, and psychology. Statistical research and development efforts have produced methods for the fitting of mixtures of distributions, estimation of modes in multivariate density functions, criteria for comparative evaluation of methods, testing for the number of clusters, evaluation of consistency of algorithms and criteria, testing goodness-of-fit of classification structures, and graphical representation and visualization.

At present, there is limited theory and formalism for directing the use of clustering and related methods of data analysis in practice. We have informal guidelines at best. There is, however, a reasonable consensus that any rational approach to applications of cluster analysis should con-

sider purpose, selection of objects, type of classification structure desired, selection and scaling of variables, choice of algorithm, criteria for evaluation, description of the meaning and appropriate communication of the proposed classification.

To successfully use cluster analysis, one must integrate the techniques of cluster analysis with the subject matter. The logic and data structures must reflect meaning and understanding in the appropriate context.

In order to accomplish this task, expert knowledge of the subject matter and problem context must be brought to bear on the key decision points in the process of any analysis that uses current clustering procedures. In the process of using cluster analysis, different methods may be used on the same dataset to provide insight and understanding into the data.

The reader is referred to Arabie, et al. (1996), Fraley and Raferty (1998), Friedman, H.P. and Rubin (1967), Banfield and Raferty (1993), Friedman, H.P. (1987), Kaufman and Rousseeuw (1990), McLachlan and Basford (1988), Mirkin (1996), and Ripley (1996). There is a Classification Society of North America that has been meeting annually since 1968 (<http://www.pitt.edu/~csna/>) and a *Journal of Classification* that is published by Springer. There is also an International Federation of Classification Societies that has been meeting since 1988 (see Gower, 1988).

Clustering methods and, in particular, model based clustering consume large amounts of computing resources. Recent interest in the clustering of massive datasets has led to the development of new algorithms and approaches to scale up existing algorithms. A good overview of work in this area can be found in the forthcoming paper by Murtagh (2000).

### 3.3 Searching for Interesting Relationships

The concept of searching for interesting relationships has its philosophical roots in artificial intelligence in the question, 'Is there a logic of discovery that is distinct from a logic of confirmation?' References to the philosophical background include Simon (1973, 1977) and Kaplan (1998). Earlier work in Bayesian econometrics includes Leamer (1978). With a stretch, one might consider Glymour, et al (1987) which uses an artificial intelligence approach to discovering causal structure. The reader can peruse the Proceedings from the *Conferences on Knowledge Discovery and Data Mining* to pursue this are further. An explicit reference for a recent conference is Agrawal, et al. (1998).

## 4.0 Statistics and Data Analysis

...all statisticians share, to a greater or less degree, a severe stress of a kind that could produce split personalities. They must work with numbers, and usually

with other symbols, that are to be manipulated according to clearly established rules; they must be happy with the sorts of (mathematical) manipulations that are probably the most secure things in human life, are, consequently, also the most rule-bound. At the other extreme, they must, at almost the same time, be honest in assessing the uncertainties of their final results. In the latter they cannot be satisfied with allowance for only the likely size of 'sampling errors', a task with which routine manipulations can often help them; they must, most particularly and responsibly, make explicit allowance for the likely size of the 'nonsampling errors'; for the extent to which the data given to them was neither what it purported to be nor what it ought to have been. No other profession must support itself over so wide a span from security to insecurity.

[Tukey, 1965, pp.23-24.]

Statistical practitioners have known for a long time that, prior to using the methods that most textbooks emphasize, there is a very important and largely neglected phase of activity which Fisher called specification and which has also been called model identification. This involves informal techniques of analysis of data, many of them graphical, aimed at looking at data in a preliminary and exploratory way in order to help understand what questions should be asked and what tentative models might be entertained. Until recent years, however, this process was regarded by the majority as not entirely respectable. Like the black art, it was widely felt that it should be conducted, if at all, only behind closed doors.

It was a stroke of genius to realize that to render 'a deed without a name' respectable, you should name it (or perhaps I should say rename it), and we are all grateful for the name 'Data Analysis.'

[Box, 1979, p.3.]

There are many definitions of statistics, but most fail to adequately capture the interplay of data, theory, context, and technology. Today, statistical practice is profoundly affected by resources for statistical computation, visual display, and access to databases.

Statisticians have not been perceived to address the informal activities of data analysis. Although these activities are largely undocumented in the statistical literature, many statisticians have been concerned with data analysis. In this section, we provide some selected references to broad views of data analysis and statistics as expressed by leading statisticians.

Much of what statisticians do in practice has not been formalized. Attempts to formalize, in an expert system, what a statistician does in building regression models have failed (Gale, 1986). It is recognized by statisticians that formal theories of statistical inference do not capture the major aspects of what statisticians do in solving problems. Savage (1981) indicates that the lack of a formal theory should not inhibit what the statistician does. For example,

while randomization has no place within the Bayesian theory of inference, Savage would recommend continued use where appropriate.

Tukey had a major influence on the theory and practice of statistics. Two volumes of his collected works (1986a, 1986b) are devoted to the philosophy and principles of data analysis. The volumes contain many important insights, guidelines, and principles for the practicing statistician.

Dempster (1981) introduces the term 'functional statistics' for integrating the technical tools of statistics with the subject. He notes that only technical statistics is taught to the statistician. Functional statistics are seen only in case studies of applications.

Mallows and Walley (1980) attempted to develop a theory of data analysis by analyzing case studies. Although they were unsuccessful, they did produce some guidelines for data analysis and identified the need for 'strategies' for approaching statistical problems in contrast to 'tactics', i.e., the choice of specific methods.

The relation of statistical theories of inference with data analysis as part of a global view of statistics is discussed in Diaconis (1985). This broad view of statistics is, unfortunately, not reflected in textbooks.

Cox (1981) describes theory and general principles of statistics; by theory, he does not mean formal mathematical theory. These principles are expanded in Cox and Snell (1982) where useful guidelines, examples, and strategies for the statistical analysis of data are provided. His recent paper (1997) describes his views of statistics.

Chatfield (1988) provides guidelines for problem solving in statistical context with worked examples in a graduate level text for a course in problem solving.

The paper by Glymour, et al. (1996) specifically addresses the data mining community and discusses some of the recent advances in statistical data analysis.

Statistics interfaces with the development of computer algorithms, machine learning, neural networks, and artificial intelligence search algorithms. In particular, Ripley, in his book, *Pattern Recognition and Neural Networks* (1996), describes the statistical foundations and performance assessments of methods developed in different areas. His attitude is exemplified by the following comment: "statistics encompasses what the community of statisticians do, of whom your author is one!" (p.3) Clearly, Ripley identifies himself as a statistician; data mining is not an entry in the index of his 1996 book.

In summary, statistics and data analysis have formal theories and informal principles and guidelines. Unfortunately, it is no surprise that many data miners who are not trained in statistics, and even some statisticians, do not have any appreciation of the potential for statistics since they have no acquaintance or awareness of the broader view of statistics that is not often taught or reflected in textbooks.

## 5.0 Opportunities for Statisticians in the Analysis of Non-Experimental Data in the Pharmaceutical Industry

Before the availability of software packages for the analysis of data, most applied statisticians found themselves writing algorithms to perform analysis. In the last 30 years, the availability of user friendly software for analysis has freed the statistician to participate more fully in problem formulation, and the interpretation and reporting of results. But, in the pharmaceutical industry today, the data from various functional areas reside in databases that may be readily accessed or may be difficult to access, and may have accumulated over time from different sources with changing conventions and structures and differing levels of quality. The development of databases and their maintenance is falling into the Informatics and Bioinformatics areas with the result that the statistician now needs to obtain access to these databases. Now, with the large amounts of data that are available for analysis or 'mining', the statistician needs to develop algorithms to access the data or to collaborate with computer scientists to develop these algorithms.

### 5.1 Drug Discovery and Preclinical Development

From discovery through development, there is a perceived need to integrate data and reports from multiple laboratories in discovery and preclinical development. Investments are being made in the development of standards for data collection and retrieval in basic science and development laboratories and for the integration of data with output from, for example, high throughput screening, genomics, toxicology, pharmacology, etc... Data warehouses are in development with data from these diverse sources. But the access for the statistician is now often more difficult because the development of these data repositories is led by the Informatics and Bioinformatics areas. The next step within these areas has been to develop algorithms for access and preprocessing and then to attempt to run analysis algorithms against the reduced data sets. Statisticians can contribute greatly to the discovery organizations through collaboration on issues of database structure, storage, and access and the identification of patterns, commonalities, and important relationships.

The issues that face the pharmaceutical industry and others can be illustrated from the current developments in the areas of the analysis of microarray data for gene expres-

sion arising from the gene chip and functional genomics data. In general, we note that these analyses are falling under the Bioinformatics rubric and are supposedly automated with the use of available software. Much of the software incorporates 'black box' preprocessing to reduce the databases to manageable sizes and then implements clustering and classification methods, generally using available software.

The analysis of gene chip microarray data is based on clustering and classification methods and methods for outlier detection. Extensive preprocessing of raw data is performed to reduce the dataset to a manageable size before the application of classification and analysis techniques. In the analysis of gene expression data, the extraction of signal from noise, for example, and the evaluation of the relationship of the observed pattern of expression to disease states require statistical collaboration. The issues that have to be addressed are described in Weir (2000) and Wong (2000). Specific statistical considerations with respect to the error structure are examined by Wittes and Friedman (1999). Lander (1999) notes that:

Computational scientists working in the field of 'data mining' have devised a dizzying assortment of techniques for clustering, predicting and visualizing patterns in high-dimensional space - most based on inherent assumptions about the types of patterns to be found. Empirical exploration will be needed to flesh out which types of datasets and analytical tools will be most fruitful for biology. (p.4).

He then raises the question of "How well can causation be inferred from correlation?" (p.4) Since much of the accumulating genomics data arises from non-experimental, attention also has to focus on the assessment of causality from a statistical perspective (cf., Freedman, 1999).

Toxicology studies that include both animal safety and toxicokinetics study as well as carcinogenicity studies are routinely carried out for all compounds that are planned for development. These studies, that incorporate placebo groups of animals, are often done in the same species using standard designs. Opportunities to study species effects in the placebo groups across multiple studies as well as to examine class effects within the same species can be exploited using statistical approaches for pooling data across studies from the 'data warehouses' that are now being created by many companies.

Similarly, the results of genetic toxicology studies can be explored across classes of compounds with appropriate pre-planning and standardization.

### 5.2 Clinical Drug Development

In the clinical drug development area, the statistician plays a leadership role in the development of data collection forms, the details of the implementation of the trials,

and standards and conventions for data processing and reporting from the trial, as well as monitoring of the conduct of the trial. Such a role places the statistician in a unique position to understand, access, and analyze the resulting data from a trial with all of its idiosyncrasies.

In the pharmaceutical industry, however, the regulatory paradigm makes it difficult for the statistician involved in clinical drug development to engage in exploratory analysis activities. Regulatory approvals require carefully planned clinical trials, the planning of which includes detailed plans for statistical analysis of the data. In this setting, the role of the statistician is often viewed as limited to the confirmatory arena. This perception of the role of the statistician as narrow and limited to the design of clinical trials to meet specific objectives by specifying hypotheses and sample sizes, testing of hypotheses, and conduct of preplanned data analyses can be frustrating for the statistician. Any analysis plan should include strategies for the examination of the underlying data, analytical assumptions, and for handling surprises. This latter situation allows the clinical trials statistician the opportunity to explore the data for understanding. Certainly, when an untoward surprise occurs, the expectations from all who are involved are that the data and results be explored and "explained". While such analyses are unlikely to yield regulatory approval, insights are obtained to further develop that compound to address the issues or to aid in the development of subsequent compounds. Cochran's work on the planning and analysis of observational studies is relevant (1982, 1983, 1985).

There are many instances in which investigators have interests that go beyond the planned purposes of the clinical trial or collections of clinical trials. Data are costly to collect and process, and, often, the primary results will make it very difficult and possibly even unethical to conduct additional trials that might shed light on etiology, treatment interactions, and subgroup effects. In addition, new hypotheses can be explored in existing datasets. Therefore, it seems reasonable to consider the use of data for purposes beyond the original intent of a regulatory submission. The statistician is in a position to contribute greatly to these activities, which are often carried out by functional groups distinct from statistics with minimal, if any, input from statistics.

Many data miners view their purpose to be the analysis of large databases to identify interesting associations and relationships in databases accumulated for other purposes. The purview of Outcomes Research has included analyses of secondary endpoints beyond the scope of the original trial, and analysis of large databases from clinical trials and marketing studies to identify interesting associations and relationships for other purposes. Many of these analyses are carried out to develop marketing claims. Since these efforts are for marketing purposes, data mining approaches have been used. There is an opportunity for statistical thought and approaches to evaluate the credibility of the claims of the data mining exercises.

Meta-analyses of clinical trials for a given indication are being carried out in increasing numbers in order to gather strength across trials. Can we generate new hypotheses for further study? Can we provide definitive answers to new questions? Can we go beyond the original goals of the individual studies to examine, for example, subgroup effects? Or, should we restrict ourselves to designing meta-analyses prospectively? Certainly those of us who are involved in meta-analysis consider that it is reasonable to do additional analysis of the existing data in new ways. Many of the issues involved in any meta-analysis (cf, Friedman and Goldberg, 1996) are statistical in nature including the choice of fixed or random effects analyses.

Many companies, today, have major strategic interests in specific therapeutic areas and/or disease indications. They have accumulated data from multiple trials, often in the same class of compound, that can be combined to provide relatively large datasets for the identification of possible new risk factors or other aspects of the diseases under study or the compounds. The opportunity to study placebo behavior across trials for the same disease entity is also available. Because the patients were accumulated in prospective trials, the opportunities to gain information from combining data that were collected under defined protocols should be exploited. cardiovascular disease, hypertension, obstructive airway disease, and cancer are just several of the areas that would be amenable to further analyses across trials. Potentially, new marketing claims could be identified through such analyses. regulatory agencies, and, in particular, the U.S. F.D.A., are in the unique position to expand these types of analyses to study placebo effects, effect sizes, and prognosis for various disease indications and compound classes. These latter data sets would likely begin to approach large sizes.

While the individual clinical trial is generally at most modest in size, although some mortality trials may include 20,000 or more patients, collections of trials begin to approach large databases and certainly contain many of the problems associated with collections of data from multiple sources.

The monitoring of accumulating safety data from clinical trials programs also often falls under the purview of safety and surveillance or pharmacovigilance organizations that are functionally separate from statistics and often from the clinical development organization as well. What is a trend with respect to the occurrence of an adverse event? What is an unexpected increase in the occurrence rate of an adverse event? Statisticians, again, are uniquely qualified to collaborate in the activities that surround the evaluation of accumulating data. They have, as part of their skills, the ability to assess the completeness of the underlying data, sources of bias, likelihood of the observed results under varying assumptions, etc. and therefore can contribute meaningfully to the ongoing dialogue regarding risk versus benefit of any product. (We are leaving aside issues involved in maintaining the blind in randomized trials that we believe can be as effectively managed with an internal

data safety monitoring board, when such a board is appropriate, as with an external board.)

The analysis of safety during the drug development process generally results in an integrated summary of safety that follows a prescribed format for regulatory approval. Much of the reported analysis is generally descriptive with statistical testing used to screen for potential problems. No adjustments for multiplicity are generally used to be conservative. The data are summarized across dosage regimens, demographic groups, types of studies. In these summaries, in the portion that summarizes the clinical trial experience, the populations at risk are known and the controlled and uncontrolled trials can be summarized. What more could be learned from such data were we to use some of the tools of exploratory data analysis, pattern recognition, etc?

### 5.3 Postmarketing Surveillance

When we enter the area of postmarketing surveillance, the data regarding adverse events generally arise from spontaneous reporting. In this case, as we are all aware, the information surrounding the report is sparse at best and subject to reporting biases and lacking meaningful denominator data. Each company engages in mandated surveillance for all of their products and regulatory agencies receive reports of serious events. How to screen these databases at the product level and at the regulatory level (for class effects) is a complex, difficult problem. The separation of signal from noise in routinely maintained databases is an area of heightened interest and activity and is an area for statisticians to contribute. Similar issues arise in the specific case of Pregnancy Registries that have been developed with the specific focus of identifying maternal and child adverse experiences associated with compounds.

The F.D.A. maintains a very large spontaneous reporting database of serious adverse events. DuMouchel (1999) used what he calls "Bayesian data mining" to screen for possible drug-adverse event associations in this database across all approved compounds. While he considers the search for interesting cells in large, sparse tables to be a data mining task, his approach is essentially statistical in nature with the objective of identifying potential increases in adverse reactions associated with a compound. The limitations of the approach are discussed in his paper and in comments by O'Neill and Szarfman, Louis and Shen, and Madigan.

Many assumptions and conventions are required to implement the several approaches to the identification of interesting cells. These assumptions themselves could influence the selection of cells. Of interest, is the comparison among the various approaches and the statistical discussions of the properties of the techniques. The influence of the conventions chosen to select the data for inclusion are not evaluated nor is any weight given to relative severity of the included serious adverse events. In particular,

small cells and infrequently occurring events would be difficult to detect with these procedures. The influence of varying the selection criteria for inclusion of cells on the results should be evaluated. Biases in the selection processes that give rise to the data in the spontaneous reporting database, along with trends in reporting, relationship of reporting "rates" to time on market and nature of the compound, etc. have potentially more serious impact on the results.

## 6.0 Summary

We agree with Cortes and Pregibon (1998) that "data mining differs from 'traditional' data analysis on one single, but important dimension, scale" (p.174). If the major difference between KDD/data mining and statistics/data analysis is the focus on "huge datasets" and the notion that analysis process can be automated, then, why not call KDD/data mining, data analysis of massive datasets? Why are the statistical software packages, SPSS, Splus, SAS, being marketed for data mining since they, at the present time, lack the capabilities to handle massive datasets regardless of machine capacity? Obviously, the vendors believe that's where the money is.

We are truly impressed that a small group of people who sponsored the first conference on KDD were able to bring together professionals from the disciplines of statistics, machine learning, artificial intelligence, and database management and exploit the hype generated in the business community by the term 'data mining'. These same few individuals started a journal in 1997, *Knowledge Discovery and Data Mining*, formed a new special interest group in the ACM (Assoc. for Computing Machinery) with a publication called SIGKDD Explorations and a website (see above).

The major issue for statisticians is that this same hype has resulted in the perception outside the statistical community that data mining is the first choice for the analysis of all datasets regardless of size. Furthermore, many supporters of the data mining viewpoint persist in viewing statistics in a very narrow context and, essentially, irrelevant to their problems.

Therefore, the real loss to business, academia, and government is the statistical expertise in the planning and process of data collection and management, planning for the integration of databases, relevance to the subject matter, awareness of pitfalls and biases arising from multiple sources of data, analysis and credibility of interpretation of results.

In addition, the information technology professionals are trying to use this hype to establish their position in the business and academic communities to be the leaders in providing data warehouses and data mining tools for analysis by all. And statistics will be used, as the drunk uses the lamppost, more for support than illumination.

And statisticians will become the mop-up squad after the fact.

For example, about 20 years ago, one of us was asked to consult on the following problem. A radiologist had data and was having no success in differentiating cells that were exposed to ultrasound from those that were not exposed. When we met, he mentioned his strong conviction that the exposed cells differed from those that were not exposed and he asked that I look at his data. Before I would look at the data, I asked for the basis of his conviction. He then showed me a motion picture of cells before and after exposure to ultrasound. The visual evidence from the film was striking. The exposed cells were moving much more slowly than the unexposed cells. He then told me that he was told by a representative of a regulatory agency that these results had to be quantified. I then asked him if the dynamics of motion were captured in his dataset. He had purchased a software package that automatically extracted features of the cells. I then looked at his data and noted that the dynamics of motion were clearly not included and suggested that he capture this information before I would attempt to collaborate with him on the development of classification rules.

The report of the Working Group on Biomedical Computing (1999) suggested that a program should be "directed toward the principles and practice of information storage, curation, analysis, and retrieval (ISCAR)". From the section on clinical trials, the group notes that "from the statistician's perspective, some problems that are labeled computational are really problems of the analysis of complex data. That analysis requires computational support, ...challenge is to create appropriate analytical tools...certainly the case with genetic array data..."

Statisticians/data analysts need to be more proactive in providing a more holistic view of their role and expertise for collaboration with subject matter experts. In addition, statisticians should be collaborating with computer scientists in the development and scale-up of algorithms for searching and analyzing large datasets as well as finding ways to bypass the need for building data vaults with the exorbitant costs and overhead to obtain the keys to accessing the contents.

### Postscript

John Tukey's obituary (New York Times, July 28, 2000) describes him at "one of the most influential statisticians of the last 50 years and a wide-ranging thinker credited with inventing the word 'software'...He was 85."

### Acknowledgements

The authors would like to thank the editors for their very helpful comments.

### Selected References

Anscombe, F.J.: *Computing in Statistical Science Through APL*. Springer-Verlag, New York, 1981.

Arabie, P., Hubert, L. J. and De Soete, G., eds: *Clustering and Classification*. World Scientific Publishing, Singapore, 1996.

Agrawal, R., Stolorz, P., and Piatetsky-Shapiro, G., eds.: *Proceedings, The Fourth International Conference on Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA, 1998.

Banfield, J. and Rafferty, A.: Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, Vol. 49, pp 803-821, 1993.

Bassett, D.E., Jr., Eisen, M.B. and Boguski, M.S. Gene Expression Informatics - It's All in Your Mine. *Nature Genetics*, 1-55, 1999.

Box, George P.: Some Problems of Statistics and Everyday Life. *Journal of the American Statistical Association*, V, 1-4, 1979.

Brachman, R.J. and Anad, T.: The Process of Knowledge Discovery in Databases: A Human-Centered Approach. In: *Advances in Knowledge Discovery and Data Mining*, pp: 37-58. Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P. and Uthurusamy, R., eds. AAAI Press/The MIT Press, Menlo Park, 1996.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J.: *Classification and Regression Trees*. Wadsworth International Group, California, 1984.

Chatfield, C.: Model Uncertainty, Data Mining, and Statistical Inference (with discussion). *Journal of the Royal Statistical Society, Series A*. 158: 419-466, 1995.

Chatfield, C.: *Problem Solving: A Statistician's Guide*. Chapman and Hall, London, 1988.

Cochran, W.G.: *Planning and Analysis of Non-Experimental Studies*. Proceedings of the Twelfth Conference on the Design of Experiments in Army Research and Testing, U.S. Army Research Office, ARO-D Report 67,2:319-336, 1967. Also in: Contributions to Statistics, W.G. Cochran, #88, Wiley, New York, 1982.

Cochran, W.G.: *Planning and Analysis of Observational Studies*. John Wiley & Sons, New York, 1983.

Cochran, W.G.: The Planning of Observational Studies of Human Populations. *Journal of the Royal Statistical Society, Series A*, Series A, 128:234-265, 1965 (with discussion). Also 85 in same book as above reference.

Cortes, C. and Pregibon, D.: Giga-Mining. pp. 174-178 In: Agrawal, R., Stolorz, P., and Piatetsky-Shapiro, G., eds.: *Proceedings, The Fourth International Conference on Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA, 1998.

Cox, D.R.: Some Remarks on the Role in Statistics of Graphical Methods. *Applied Statistics*, 27:4-9, 1978.

Cox, D.R.: Theory and General Principle in Statistics. *Journal of the Royal Statistical Society, Series A*, 144. 3:289-297, 1981.

Cox, D.R. and Snell, E.J.: *Applied Statistics: Principles and Examples*. Chapman and Hall, London, 1982.

Cox, D.R.: Some General Aspects of the Theory of Statistics. *International Statistical Review*, 54:117-126, 1986.

Cox, D.R.: Discussion following Chatfield, C.: Model Uncertainty, Data Mining, and Statistical Inference (with discussion). *Journal of the Royal Statistical Society, Series A*, 158: 419-466, 1995.

Cox, D.R.: The Current Position of Statistics: A Personal View. *International Statistical Review*, 65:261-290, 1997.

Dempster, A.P.: An Overview of Multivariate Data Analysis. *Journal of Multivariate Analysis*. 1:316-346, 1971.

Dempster, A.: Purposes and Limitations of Data Analysis, Research Report S-85, Department of Statistics, Harvard University, 1981.

Diaconis, P.: Theories of Data Analysis: From Magical Thinking Through Classical Statistics. In: *Exploring Data Tables, Trends, and Shapes*, pp: 1-36, Hoaglin, D.C., F. Mosteller, Tukey, J.W. eds. John Wiley & Sons, New York, 1985.

DuMouchel, W.: Bayesian Data Mining in Large Frequency Tables, With and Application to the FDA Spontaneous Reporting System. With discussion. *The American Statistician*, 53:177-202, 1999.

Elder IV, J. and Pregibon, D.: A Statistical Perspective on Knowledge Discovery in Databases in Advances. In: *Knowledge Discovery and Data Mining*, pp: 83-116, Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P. and Uthurusamy, R., eds. AAAI Press/The MIT Press, Menlo Park, 1996.

Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P.: From Data Mining to Knowledge Discovery: An Overview. In: *Advances in Knowledge Discovery and Data Mining*, pp: 1-36, Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P. and Uthurusamy, R., eds. AAAI Press/The MIT Press, Menlo Park, 1996.

Feynman, R.P.: *The Meaning of It All: Thoughts of a Citizen Scientist*. Perseus Books, Reading, 1998.

Fraley, C. and Raftery, A.E.: How Many Clusters? Which Clustering Method? Answers via Model-Based cluster Analysis. *Computer Journal*. 4, 1: 578-588, 1998.

Freedman, D.: Some Issues in the Foundations of Statistics. *Foundations of Science*, pp. 19-39, 1995.

Freedman, D.: From Association to Causation: Some Remarks on the History of Statistics. *Statistical Science*, 14:243-258, 1999.

Friedman, H.P. and Goldberg, J.D.: Meta-Analysis: An Introduction and Point of View. *Hepatology*, 23:917-928, 1996.

Friedman, H.P. and Rubin, J.: On Some Invariant Criteria for Grouping Data. *Journal of the American Statistical Association*, 62:320:1159-1178, 1967.

Friedman, H.P.: Strategies for Multivariate Data Analysis: Case Study. In: *Lecture Notes in Medical Informatics: Acquisition, Analysis; and Use of Clinical Transplant Data*, JanBen. R. and Opelz, G., eds., pp.: 51-68. Springer-Verlag, Berlin, 1987. vol. 34, 1987.

Friedman, J., Hastie, T. and Tibshirani, R.: Additive Logistic Regression: A Statistical View of Boosting. Second revision, Stanford Technical Report, 1999.

Gale, W. and Pregibon, D.: Artificial Intelligence Research in Statistics. *The AI Magazine*, pp. 72-75, Winter, 1985.

Gale, W.: *Artificial Intelligence and Statistics*. Addison Wesley Press, 1986.

Gifi, A.: *Nonlinear Multivariate Analysis*. John Wiley and Sons, Inc. Chichester, 1990.

Glymour, C., Madigan, D., Pregibon, D., Smyth, P.: Statistical Inference and Data Mining. *Communications of the ACM*, 39, 35-41, 1996.

Glymour, C., Madigan, D., Pregibon, D., Smyth, P.: Statistical Themes and Lessons for Data Mining. *Data Mining and Knowledge Discovery*, 1:11-28, 1996.

Glymour, C., Scheines, R., Spirtes, P. and Kelly, K.: *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*. Academic Press, Orlando; 1987.

- Goebel, M. and Gruenwald, L.: A Survey of Data Mining and Knowledge Discovery Software Tools. *SIGKDD Explorations*, 1:20-33, 1999.
- Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286:531-538, 1999.
- Gorenstein, S.: Database Reorganization via Clustering. *Proceedings of COMPSTAT International Symposium on Computational Statistics*, Sept., 1974, Vienna, Austria, Physica-Verlag.
- Gower, J.C.: *Classification, Geometry, and Data Analysis*. In: *Classification and Related Methods of Data Analysis*, pp: 3-14, Bock, H.H., ed. North-Holland, Amsterdam, 1988.
- Hand, D.J.: *Construction and Assessment of Classification Rules*. John Wiley and Sons, Chichester, 1997.
- Hand, D.J.: Data Mining: Statistics and More? *The American Statistician*, 52:112-118, 1998.
- Hand, D.J., Manilla, H and Smyth, P.: *Principles of Data Mining*. Cambridge, MIT Press, (forthcoming).
- Hand, D.J.: Statistics and Data Mining: Intersecting Disciplines. *SIGKDD Explorations*, ACM SIGDD, V.1, 16-19, 1999.
- Holland, J.H., Holyoad, K.J., Nisbett, R.E. and Thagard, P.R., eds.: *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, the MIT Press, 1980.
- Huber, P.J.: From Large to Huge: A Statistician's Reactions to KDD & DM. *Proceedings Third International Conference on Knowledge Discovery and Data Mining*, pp: 304-308, Heckerman, D., Mannila, H. Pregibon, D. and Uthurusamy, R., eds., 1997.
- Jensen, D.: Data Snooping, Dredging and Fishing: The Dark Side of Data Mining: A SIGDD99 Panel Report, *SIGDD Explorations*, ACM SIGDD, V.1, 52-54, 2000.
- Kaplan, A.: *The Conduct of Inquiry: Methodology for Behavioral Science*. Transaction Publishers. New Brunswick, 1998.
- Kaufman, L. and Rousseeuw, P.J.: *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Inc., New York, 1990.
- Lander, E.S.: Array of Hope. *Nature Genetics Supplement*, 21:3-4, 1999.
- Leamer, E. E.: *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. John Wiley and Sons, New York, 1978.
- Lovell, M.: *Encyclopedia of Statistics*. Update, Vol. 3, 1999
- Michalski, R.S., Carbonell, J.G. and Mitchell, T.M., eds.: *Machine Learning: An Artificial Intelligence Approach*. Tioga Publishing Company. Palo Alto, 1983.
- Mallows, C.L., and Walley, P., eds.: A Theory of Data Analysis? *Proceedings of the American Statistical Association, Business and Economics Section*. pp: 8-14, 1980.
- McLachlan, G.J. and Basford, K.E.: *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, Inc., New York and Basel, 1988.
- McLachlan, G.J.: *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, Inc., New York, 1992
- Mirkin, B.: *Mathematical Classification and Clustering*. Kluwer Academic Publishers, Dordrecht, 1996.
- Moses, L.E.: Statistical Concepts Fundamental to Investigations. *The New England Journal of Medicine*, 312:890-897, 1985.
- Murtagh, F.: Clustering in Massive Data Sets. In *Handbook of Massive Data Sets*, eds. Abello, J., Pardalos, P.M., and Reisende, M.G.C., Kluwer, 2000 (forthcoming).
- Piatetsky-Shapiro, G.: Knowledge Discovery in Databases: 10 Years After, *SIGKDD Explorations*, ACM SIGKDD, V.1, 59-61, 2000.
- Rao, C.R.: The Use and Interpretation of Principal Components in Applied Research. *Sankhya*, Series A, 1964.
- Ripley, B.D.: *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

Savage, L.J.: *Conference on the Future of Statistics: Proceedings*, pp: 146. D. Watts, ed. Academic Press, New York, 1968.

Savage, L.J.: *The Shifting Foundations of Statistics in The Writings of Leonard Jimmie Savage: A Memorial Selection*, pp: 721-736. American Statistical Association and The Institute of Mathematical Statistics, 1981.

SIGKDD: *Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, Association for Computing Machinery*. 1:1, 1999.

Simon, H.: *Models of Discovery*. D. Reidel, Dordrecht, Holland, 1977.

Simon, H. A.: Does Scientific Discovery Have A Logic? *Phil. Sci.*, 40, 471-480, 1973.

Speed, T. and Waterman, M.S., eds.: *Genetic Mapping and DNA Sequencing*. Springer-Verlag, New York, 1996.

Tukey, J.W.: The Technical Tools of Statistics. *American Statistician*, 19:23-28.

Tukey, J.W.: *The Collected Works of John W. Tukey: vol. III, Philosophy and Principles of Data Analysis: 1949 - 1964*. Jones, Lyle V., ed. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA, 1986(a).

Tukey, J.W.: *The Collected Works of John W. Tukey: vol. IV, Philosophy and Principles of Data Analysis: 1965 - 1986*. Jones, Lyle V., ed. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA, 1986(b).

Venables W.N. and Ripley, B.D.: *Modern Applied Statistics with S-PLUS*. Third edition. Springer-Verlag, New York, 1999.

Weir, B.S.: Challenges Facing Statistical Genetics. *Journal of the American Statistical Association*, 95:319-322, 2000.

Wittes, J. and Friedman, H.P.: Searching for Evidence of Altered Gene Expression: a Comment on Statistical Analysis of Microarray Data. *Journal of the National Cancer Institute*, 91 400-401, 1999.

Wong, W.H.: Computational Molecular Biology. *Journal of the American Statistical Association*, 95:322-326.

Working Group on Biomedical Computing. Advisory Committee to the Director, National Institutes of Health, June 3, 1999: The Biomedical Information Science and Technology Initiative.

## Letter from the Chair

### Thomas Capizzi

I believe that I will get one more crack at this column when as outgoing chair, I will describe more fully the state of the section. For now, I'll content myself with summarizing some recent section activities.

### 2000 FDA-Industry Workshop

I am writing this column a few days after the 2000 FDA-Industry Workshop on Statistically Sound Decision Making at the Bethesda Md. Hyatt-Regency. This workshop which is cosponsored by the section and the FDA Statistical Association was a resounding success with a record attendance of over 360 attendees. This probably means that the conference will probably lose more money than last year as a result of the labyrinthine ASA revenue sharing rules..

Nearly all of the credit for this conference goes to the co-chairs, Greg Campbell and Sandy Heft, and the sixteen individuals that they recruited to serve on the program committee. The workshop committee planned 11 highly relevant sessions that featured presentations from government, academic and industry speakers on topics covering statistical policy, methodology, applications and consulting skills. The remaining credit goes to the ASA meetings department, in particular Elaine Powell and Linda Minor, for their excellent local arrangements and flawless registration. I hope that the attendees feel that they made a sound decision in attending. I look forward to next year's meeting which is tentatively scheduled for the first week in October at the same location.

In my view, this workshop has become the premier model for Industry-Government interaction and others are taking notice. Greg and Sandy have been recently asked to submit a number of papers from the workshop for publication in the Drug Information Journal. They are currently working with the presenters to prepare papers for submission So if you could not attend, you will have the opportunity to read some of the papers.

### JSM 2000

Speaking of presentations, all of our JSM sessions, roundtable luncheons, and sponsored short courses were well-received. Nearly all of the sessions that I attended were standing room only which clearly indicates the great interest that our sessions generate. Thanks go to Bob Small for organizing the sessions and short courses and to Lucas Makris and Keith Soper for organizing the roundtable luncheons. I also want to thank Doug Faries and Viswanath Devanarayan for coordinating the Best Contributed Paper Presentation Competition.

Our JSM mixer and business meeting was also well attended with well over 100 attendees who took this

opportunity to socialize, learn about section activities, and to congratulate our 1999 Best Presentation and 2000 Best Student Paper winners (see related articles).

## Newly Elected Officers

Congratulations to our newly elected officers for 2001: Bob Small, Chair Elect; Tuli Cnaan, Council of Sections Representative; Demissie Alemayehu, Publications Officer; and Len Oppenheimer, Program Chair-elect. Len is already planning for the 2002 scientific program so please contact him with your suggestions

Finally, if you have not already done so, I encourage you to get actively involved in our section's activities. You can get involved in a wide range of activities. Some examples include serving on one of our standing committees like membership, joining the FDA-Industry workshop program committee, chairing a JSM contributed paper session, etc.. If you are willing to participate more formally in our activities, please contact Jeff Meeker, the 2001 chair or myself to get on our volunteer list. If we can't find a place for you now, you can rest assured that you will remain on our list until we can find a mutually satisfactory appointment.

## Section News

# Minutes of ASA Biopharmaceutical Section

## Executive Committee Meeting August 15, 2000

*Attendees: Tom Bradstreet, Greg Campbell, Katherine Monti, Bob Small, Sally Greenberg, Jeff Meeker, Tom Capizzi, Keith Soper, Demissie Alemayehu, Sandy Heft, Ram Suresh, Len Oppenheimer, Bruce Binkowitz, Steve Snappinn, Avital Cnaan.*

The invited & contributed sessions, short courses, and luncheon roundtables were all well received. We were granted an all time high of four invited sessions. The topics of the invited sessions were: *Methods of Time to Recurrent Events Analysis*, *Topics in Health Economics*, *Methodological issues in Meta-Analysis*, and, *What's new in the Statistical Evaluation of Pulmonary Drugs*. We sponsored three Topic Contributed Sessions and had a total of seven Regular Contributed sessions. Two short courses, *Statistical Methods in Modern Molecular Biology* given by E.Lazaridis, P.Gieser & G.Yanev, and, *Equivalence Trials: Statistical Issues* given by K.Ghosh, N.Biswas & I.Chan were also sponsored.

For the 1999 ASA Biopharmaceutical Section best contributed paper awards, R.A. Railkar (first place), N.R.Bohidar (second place) and L.Helms (third place) were

announced as the winners. All the papers submitted for the 2000 ASA biopharmaceutical student paper competition were of very good quality. Five participants, J.Bryan (Univ. California, Berkeley), Z.Shang (North Carolina State Univ.), V.Somayaji (Pfizer), D.Yu (Univ. of Michigan) and Q.Yu (Quintiles), were chosen for the awards.

Sandy Heft and Keith Soper reported progress to date for the 2001 program. The Biopharmaceutical section has been given 4 invited sessions for JSM 2001. Eight potential topics have been received by Soper, and four will be chosen after consultation with Biometrics, ENAR and WNAR in order to avoid overlap. Input on these topics from the executive committee members will also be used in the determination of the final four topics. No proposals for short courses have been received yet. At the 2001 ENAR meetings, two sessions will be sponsored by the Biopharmaceutical section. In addition, we will be the cosponsor of three sessions.

The ASA proposal to produce a searchable CD ROM that will include proceedings from all sections was endorsed by the executive committee. The committee also suggested an alternative price to that suggested by the ASA.

The last issue of the Biopharmaceutical Report was well received. The Fall issue to come out in September and the last issue of 2000 to come out in December.

The web site is up to date as of August 15,2000. The only part that is not updated is the mailing list. Need to acquire membership list from ASA.

Bob Davis (Chair 2001), Bob Starbuck (2001-2002), Christie Chung-Stein (2001-2003) have been appointed to serve in the ASA Fellows committee.

The Biopharmaceutical Section has four offices up for election in 2001: Section Chair for the year 2003 (Chair-Elect in 2002 and Past-Chair in 2004), Program Chair for the year 2003 (Program Chair-Elect in 2002), Secretary-Treasurer for the years 2002-2004, Council of Sections Representative for the years 2002-2004. Recommendations for candidates have come from the ASA office and directly from the membership (one apiece), and these potential candidates were carefully considered. A proposed slate of candidates will be finalized after discussions with the current and future Section Chairs.

Sally Greenberg presented the budgeted versus estimated revenue/expenses as of 6/30/2000. We are on target at this point with respect to revenue.

Bruce Binkowitz of the membership committee updated the executive committee on the priority of needs. The first priority will be the development of a new survey of members as an update to that done in 1996. The committee agreed with this, and suggested we look into using the internet as well making the survey shorter than the previous one. The second priority concerns determining the value, and creating incentives, for why an individual would want to join the section, and also why a corporation would want to sponsor the section. Discussion then ensued on the difficulty of getting an electronic listing of all members from ASA. Tom Capizzi will work with the ASA office in an attempt to resolve this issue.

Tom Capizzi discussed the revisions that he/Jeff Meeker made to the Manual of Operations. In general, the execu-

tive committee generally supported the revisions and asked for clarity with respect to the program chair responsibilities.

Jeff Meeker announced the following appointments:

- ◆ **Executive Committee (2001-2003):**  
Anne Cross, Tom Bradstreet
- ◆ **Associate Editor, *Biopharmaceutical Report***  
Neal Thomas
- ◆ **Fellows Committee**  
Bob Davis (Chair 2001)  
Bob Starbuck (2001-2002)  
Christie Chung-Stein (2001-2003)
- ◆ **Student Paper Committee**  
Tom Bradstreet (chair),  
Sanat Sarkar, Christie Clark
- ◆ **Webmaster**  
Kalyan Ghosh
- ◆ **Mail list Coordinator**  
Sally Greenberg.

Note: Tom Capizzi will schedule a transition meeting in the month of October. More detailed minutes will be posted on the section's web page after Executive Committee review and approval during this meeting

---

## 2001 ASA Election Candidates

### Steve Snapinn

The Nominations Committee announces the following-candidates for Biopharmaceutical Section Offices in the 2001 election.

- ◆ **Chair-Elect:**  
Samuel M Heft, Schering-Plough Research Inst.  
Nancy D. Smith, Food and Drug Administration
- ◆ **Program Chair-Elect:**  
Stacy R. David, Lilly Corporate Center  
Mani Y Lakshminarayanan, Centocor, Inc.
- ◆ **Secretary/Treasurer:**  
Kalyanbrata Ghosh, Merck & Company, Inc.  
Lukas Makris, BioCor, L.L.C.
- ◆ **Council of Sections Representative:**  
Kay Larholt, Pharmaceutical Research Inst,  
Johnson & Johnson  
Naitee Ting, Pfizer Global Research & Development

## Joint Statistical Meetings Best Contributed Paper Awards

### Denise J. Roe

At the Biopharmaceutical Section Business Meeting and Mixer held during the Joint Statistical Meetings, the winners of the 1999 Best Contributed Paper competition and the 2000 Student Paper competition were announced. The Best Contributed Paper Award was established to encourage quality presentations in section sponsored contributed sessions. The 1999 award winners were based on ballots distributed during the 1999 JSM in Baltimore, Maryland. The winners and their papers are:

**First Place:** Radha A. Railkar, Temple University, "A Simultaneous Testing Strategy for Comparing Two Treatments in a Stratified Binomial Trial" Co-authors: Devan V. Mehrotra and Boris Iglewicz.

**Second Place:** Norman R. Bohidar, Merck Research Labs, "A Comparison of Approaches for Analyzing Longitudinal Data when the Range of Baseline Scores is Restricted" Co-authors: Duane Snaveley, Joseph Pigeon, and Zhongxin Zhang.

**Third Place:** Laura Helms, University of North Carolina Chapel Hill, "A Comparison of Multiple Imputation and Traditional Missing Data Methods in Clinical Trials" Co-author: C.E. Davis.

The goal of the Student Paper Competition is 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 JSM. The 2000 award winners were selected in a blinded manner by a committee of section members (Avital Cnaan (chair), Thomas Bradstreet, Christine Clark, and Sanat Sarkar). The five winners and their papers are:

**Jennifer Bryan**, University of California Berkeley, "Gene Expression Analysis with the Parametric Bootstrap" Advisor: Mark J. van der Laan, University of California Berkeley

**Zhe Shang**, Agouron Pharmaceuticals, Inc., "Expected Survival Based on the Proportional Hazards Model" Advisor: Anastasios Tsiatis, North Carolina State University

**Veena Somayaji**, Pfizer, Inc., "Marginal Analysis of Multivariate Mixed Discrete and Continuous Responses with Clustering" Advisor: Mark Becker, University of Michigan

**Daohai Yu**, University of Michigan, "Analysis of Interval-censored Time-to-event Data with a Marked Observation Process" Advisor: Robert Wolfe, University of Michigan

**Qifeng Yu**, Quintiles, Inc., "Parametric Bootstrap Based Inference in Linear Mixed Effects Model with AR (1) Correlated Data" Advisor: Shie-Shien Yang, Kansas State University

Congratulations to each of the winners. For more information about these competitions please contact the Biopharmaceutical Section through <http://www.amstat.org> or directly at <http://www.best.com/~asabpl/>.

Note: The above was also submitted to the *Amstat News*

---

## The 24th Annual Midwest Biopharmaceutical Statistics Workshop

### James Nezamis

The 24th Annual Midwest Biopharmaceutical Statistics Workshop will be held on the campus of Ball State University in Muncie, Indiana, on May 21-23, 2001. This workshop is co-sponsored by the Biopharmaceutical Section of the American Statistical Association. A preliminary program will be available in December, 2000. For more information contact:

**James Nezamis, Publicity Chair**

Quintiles Inc.  
e-mail [james.nezamis@quintiles.com](mailto:james.nezamis@quintiles.com)  
phone: (816) 767-3849

**Ken Gerald, Workshop Chair**

Applied Logic Associates Inc.  
e-mail [kgerald@alogic.com](mailto:kgerald@alogic.com)  
phone: (713) 529-4747 ext 170

**Mir Ali, Local Arrangements Chair**

Ball State University  
e-mail [mali@wp.bsu.edu](mailto:mali@wp.bsu.edu)  
phone: (765) 285-8670.

## 2000 ASA Fellows

### A. Lawrence Gould

*The following members of the Biopharmaceutical Section have been elected fellow of the ASA:*

**Paula K. Roberson**, Professor and Director, Division of Biometry, University of Arkansas: For excellence as a statistical collaborator in clinical research; for fostering of statistical consulting and education; and for service to the profession.

**Frank W. Rockhold**, Vice President and Director, Biostatistics and Data Sciences, SmithKline Beecham Pharmaceuticals R&D: For excellence in leadership in the application and promotion of statistics in the pharmaceutical industry; for administration and development of statisticians; and for promoting industrial support of academic programs in statistics.

**Edward F. Vonesh, Jr.**, Technical Director of Biometrics, Baxter Healthcare Corporation: For contributions to the theory and application of models for the analysis of repeated measurements, especially nonlinear mixed-effects models; for excellence in collaborative research in health care; and for service to the profession.

*Congratulations to all.*

Our 2001 fellows committee will be chaired by Bob Davis who will be assisted by Christy Chuang-Stein and Bob Starbuck. Their mission is to facilitate the fellows nomination process for deserving section members. To this, they need your support in suggesting potential fellows and in helping out with nominations packages. So please contact them with your recommendations or comments.

## Let's Hear from You!

If you have any comments or contributions, contact Editor Demissie Alemayehu, Biometrics Director, 235 East 42nd Street, Bldg 205/4, Pfizer Inc, New York, NY 10017; e-mail: [alemad@pfizer.com](mailto:alemad@pfizer.com); Associate Editor Kannan Natarajan, Bristol Myers Squibb, P.O.Box 5400, Princeton, NJ 08543; Phone: 609-818-4299; Fax: 609-818-5740; e-mail: [Kannan.Natarajan@bms.com](mailto:Kannan.Natarajan@bms.com) ; or Past Editor Ersen Arseven, Arseven Consulting, Inc., 55 Old Nyack Turnpike., Suite 606, Nanuet, NY 10954; Phone: 845-627-1321; e-mail: [earseven@spyrat.net](mailto:earseven@spyrat.net).

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

(c) 2000 *The American Statistical Association*  
 Printed in the United States of America

NON-PROFIT ORG  
 U.S. POSTAGE  
**PAID**  
 ALEXANDRIA, VA  
 PERMIT NO. 361

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

