

# Comparing R&D and Commercial: Are All Data Created Equal?

Twenty-fifth Annual Midwest Biopharmaceutical Statistics Workshop  
Muncie, Indiana  
21 May 2002

Rafe M. J. Donahue, Ph.D.  
Principal Consultant Data Mining  
GlaxoSmithKline  
Research Triangle Park, North Carolina  
rd34115@gsk.com  
<http://home.earthlink.net/~rafedonahue>

## Some background

In the early summer of 2001, I made a move from the safe and friendly confines of Medical Affairs Statistics at GlaxoSmithKline (GSK) to the unknown world of Data Mining Technologies, a part of our Commercial IT department. Since arriving at GSK from Hoechst Marion Roussel (HMR) and its predecessor, Marion Merrell Dow (MMD), in 1996, I had been involved in the various statistical components of Phase IV clinical trials. While this certainly was a change from the Phase I, II, III world I had experienced at HMR and MMD since 1992, it wasn't vitally different. All the classic statistical elements of clinical trials were there: protocols, analysis plans, sample size calculations, primary analysis reports, secondary and exploratory analyses, patient listings and summaries, project development plans, and the like.

There were, however, certain things about Phase IV that were different. The general feel was slightly dissimilar since the focus of the development plan was rarely a pair of adequate and well-controlled trials each with  $p$ -value less than 0.05. Thus, there was more latitude in study design. Principles of sound statistical and scientific conduct were still essential but the foci of the development plans were varied. Prior to a regulatory submission in Phases I, II, and III, all activity converges toward a single point: the submission. After approval in Phase IV, though, the program is more free to expand to other areas and is no longer focused solely on "the submission."

But the move from pre-submission work to post-submission work was nothing compared to the change from clinical trials to the world of Commercial.

In the year that I have spent exploring this brave new world of commercial data, I have come to see differences and similarities and opportunities for learnings that are available to the statisticians working with both clinical trial and commercial data.

## An apology

My intent when I began this paper was to come up with some cute little mnemonic for remembering the differences between R&D and Commercial or to build little categories of the types of differences but this proved to be harder than I anticipated. Furthermore, as I began to build this list, I found that there were opportunities for learnings on both sides of the fence. In seeing another area of the company that relies heavily on data, I was able to see where each side was doing well and doing poorly — and there is often overlap on both the good and bad sides. Therefore, what follows are three general areas of learnings and the lessons that I have learned in being able to see two different areas of the company. I'll attempt to address these areas where each camp can learn from the other and what I have learned from comparing the two camps.

## Primary versus secondary uses of data

In my pocket is my keychain. On this keychain, along with a key to my car, a key to my house, a key to my office, a key to the basketball room in the gym at the church, and a key I no longer remember which lock it fits, are three small plastic cards that (1) identify me as an MVP Customer at Food Lion, (2) get me S&H greenpoints (whatever they are) at Lowes Foods, and (3) just generally save me money at Kroger. When I go through the checkout line at these businesses, I scan my card and get savings that the non-card-bearing public doesn't get.

The cash register was invented in 1884 by James Ritty and improved by John Patterson and The National Cash Register Company as a way to record transactions to keep the moneymakers from pocketing the profits. The cash register essentially records details of the transaction (the data) so as to monitor how much money ought to be in the drawer at the end of the shift.

That's the primary use for those data.

But the items you buy at the store aren't scanned just to make sure the right price is entered and to speed you along. Home Depot and Wal-mart pioneered the use of the register data for inventory control. Thus, when enough boxes of Tide are purchased and recorded in the system, the central distribution center can disburse enough replacements so as to allow you to keep your clothes clean. And it saves Wal-mart money since they don't have to keep people wandering the aisles to keep track of inventory.

Of course, the system isn't perfect, like when the checker last week couldn't find the right scanning number for a lemon and two grapefruit that I purchased and instead charged me for one grapefruit and two packages of yeast. But the system uses the transaction data for a secondary use that helps its business. How helpful is it? Who is the leader in home improvement? And who is the world's largest retailer with 220 billion dollars in sales?

So, why do I need to use the little plastic cards to get bonus savings on purchases when I present the card? In exchange for giving out my name and address to my local food retailer, I get savings and *the retailer* gets information about what kinds of things customer number 429215160345 likes to buy.

Hence, for a small loss in savings given to the customer, the company gets information about the customer. And everytime I come into the store and use my card, they find out more about me.

These are secondary uses of those data which have a primary use of documenting a transaction.

When I go to the doctor's office and pick up my allergy prescription, no data are collected. Only when I go to the pharmacy and turn in my prescription and my insurance information, are there data collected. The data document my transaction. But its secondary uses drive nearly all the statistical work that takes place in Commercial.

On the other side of the coin, in R&D, the data in a clinical trial are typically design for precisely the use for which they are used. The data come from a *designed experiment*. That is, the study is executed via a specific data collection strategy. That strategy is constructed and detailed in the protocol and analysis plan to allow the eventual analyses to be the *primary* use of the data.

A question is posed in a clinical trial, a method is designed to answer that question, and data are collected to fulfill the method. The primary use of the data is answering the question.

Contrasting this with the analyses in Commercial leads to the first major difference between the two camps. Commercial statisticians have the lot of working mostly with data that have a *different* primary use than the one for which they are using it.

Does this mean that the secondary uses of the data are worthless? Are analyses based on these data invalid? Not necessarily. The analyses are certainly not up to the same rigor as those that make primary use of the data since they are not from designed experiments. They can, however, be used to make important decisions. The automatic coupon dispenser at the checkout stand typically gives me coupons for Poptarts. Guess what Dad typically purchases for his kids on his short, evening trips to the grocery store? Yep, milk, cereal, and Poptarts.

### **Lesson one: Understand *why* your data were collected.**

#### **Analysis versus reporting**

I thought I would look up the word *analysis* in the dictionary since it is used so often in our field. People are always "doing an analysis." The word *analysis* seems to come from a Greek work *analyein* which means "to break up" or "to loosen" or "break into its component parts." Is this what we are doing when we do our analyses?

In R&D, certainly in the context of clinical trials, we do lots of analysis, I think. Surely the primary analysis of a clinical trial seems to be "analysis" in that we break the data down to the fundamental question of "Did the drug beat placebo?" or "Did we show equivalence?". But if we are really doing analysis, we probably need to be certain to include our secondary investigations in our "analysis" of the data. If we want to look at the component parts of the data from a trial, we most certainly need to examine secondary endpoints as well. And after we are finished doing what we planned to do in the protocol and analysis plan we probably need dig around in the data even further, doing *ad hoc* investigations to really distill and dissolve the information that are in the data from the trial.

What happens on the Commercial side? Does analysis take place? The data on sales of our products and our competitors' are devoured and distilled by district and territory and zip code and market segment and product strength and every possibly-imaginable split and level. But typically the data are simply aggregated and *reported*; analysis, as we might expect in R&D, seems to me to be rather rare. It is certainly done sometimes but the majority of the statistical effort seems to be focused on aggregation and reporting.

Data reporting is also done in R&D. Certainly in clinical trials, we generate patient listings of all the raw data from the trial. We also produce lists of patients who exhibit certain characteristics (those who withdraw early, those who died, etc.), lists of common adverse experiences and concomitant medications, and patient narratives.

So, is there a difference between analysis and reporting? I think that there is and that the difference, oddly enough, has to do with the dimensionality of the data.

All the data we collect are multivariate. In clinical trials, the experimental units are patients; each patient generates a vector of data. On the Commercial side the experimental units are harder to define. They may be prescribers if our interest is in understanding different physicians' prescribing habits. They may be payment organizations (hospitals, insurance companies, HMOs, etc.) if our interest is in understanding payment plans. They may be pharmacies, if we seek to understand pricing and coupon plans. They may again be patients if we seek to look at drug compliance and interactions.

Regardless of the experimental units, however, the data are multivariate in nature and, as such, the *relationships between the elements* of the data vector are at least as important as the behavior or distributional characteristics of the individual elements of the data vector.

Thus, in order to do *analysis* of data, whether those data are employed in a primary or secondary (or even tertiary!) use, we need to do more than just report on the univariate elements of the data vectors; we need to understand the multivariate nature of the data — and this means that we need to do more than just reporting.

I'll say then that analysis begins when we start to look at at least the bivariate relationships between variables. This might be simple cross-tabulations or bivariate frequency counts but to get to the component parts of data, we need to understand the multivariate relationships. As such, graphical methods for displaying more than one variable at a time are necessary for analysis as are investigations of interactions and examinations of more than one endpoint.

The tables and figures that we generate need to be designed to facilitate the comparison of interest. Even the reports that we produce often don't allow us to discover the underlying secrets in the data. The development of tables and figures to display our data is a vital component of analysis planning in both R&D and Commercial and should receive as much attention as selecting the right statistical methods and summary statistics for the analyses. Only if we are familiar with the questions that will be answered with the graphics and reports and tables and listings that we produce will we be able to produce output that answer those questions. Seemingly insipid concepts such as the sort order of data in listings can make a difference in understanding the underlying data. Often the same listing presented in two different sort orders can tell two different stories about the data. Knowing the use of the output allows the statistician to design more useful output.

**Lesson two: Data are, by nature, multivariate. Depict this nature in reports and analyses.**

### **Craftsmanship versus production**

Those who have worked with me over the past decade have become accustomed to my incessant rantings about the wasteful nature of the programming that is done at the conclusion of each clinical trial. Each trial, it seems, is treated as different and unique unto itself. I have longed for the days when nightly or weekly, blinded adverse event reports and enrollment updates and patient progress reports were generated while I slept and emailed to my inbox for my perusal the next day. The clinical trials computing system, I have exclaimed, needs to be more automated. My investment accounts (along with those belonging to millions of others) are updated nightly based on data from financial markets around the world. I can jump onto the web and watch my portfolio lose its value from essentially anywhere in the world. Why can't the clinical trial system do the same sort of thing?

It seems that each trial has its database set up just differently enough to require separate programming to produce the same kinds of tables and reports and figures that numerous other studies have produced. Why can't we just hit a button and have these standard reports produced automatically?

We have a craftsman's mentality of hand-processing each trial to get a perfect (and essentially unique!) collection of output for each trial. Yes, we are getting highly customized and sophisticated reports and analyses but why can't we produce 90% of the output from a production mindset where the data are processed according to a standard? How much resource is wasted on reinventing the wheel for each protocol?

It was precisely the production mindset that was missing for the land of clinical trials that was the first thing I noticed when I moved to Commercial IT. These guys are production oriented!

Monthly data on sales and marketing information comprising terabytes of data are processed monthly using a system so complex that no one could understand the whole thing. The sales force reps receive updates automatically in their inboxes. Regional managers get their reports automatically downloaded when they log in. The system is updated nightly with data collected from thousands detail reps making over 50,000 contacts with prescribers every day! This places swims in production! Nothing is done *ad hoc*. Everything is slick and quick and polished!

Yet, things aren't as rosy as they might appear. The system that appeared to be so fancy and automated on the surface was laced with inefficiencies and rate-determining manual steps. The process wasn't quality-checked before the data went in; only after the output reports and tables were completed several days later were things reviewed. Entire columns of zeroes were the first sign that something was amiss. (Earlier quality checking has recently been added to the system.)

*Ad hoc* reports (and analyses?) also have a place in the system. If a district manager wants a new report on activity in the district, programmers work to construct such a report. Of course, no statisticians are consulted to assist in development of these reports. And eventually the new report is added to the monthly queue with all the others so the process only gets bigger. As a result, we now generate nearly 10,000 monthly reports. (Of course, this is a source of preverse pride for some of the data warehouse people.)

The data, after production, are stored in a collection of databases (the "data warehouse") so as to allow analysts to do additional analyses that won't necessarily turn into monthly reports. The warehouse, however, is designed for efficiency of *storage*, not efficiency of analysis. As you might expect, this makes further forays into the data frightening complex and technically daunting.

How do we bring both sides toward the middle? How can we make R&D more production oriented to improve efficiency? How can we make Commercial more craftsmanlike to be able to handle intelligent *ad hoc* analysis requests and do more analysis in general?

First, statisticians in R&D need to relinquish more control of the clinical trials data to the computers/database/IT people and statisticians in Commercial need to assert more control of the design of Commercial analyses and reports from the computers/database/IT people. Statisticians add value through design of analyses and reports; IT adds value through efficient production of analyses and reports. To have statisticians doing IT-type work (as in R&D) and having IT-types doing statistics (as in Commercial) is problematic.

Second, statisticians need to focus on designing analyses and reports that get to the "component parts" of the data. This becomes then not a strictly technical role of programming and calculating but one of consulting so as to facilitate the understanding of the essence of the data. We need to cede control of the data and encourage exploration and understanding. We need to share our understanding of analysis to allow the end users to understand the data. We need to open our doors, come out of our offices, and take to users our understanding of data. We cannot think of ourselves, or allow ourselves to be thought of, as simply data clerks. The businesses demand more from us than that.

And we need to let the IT-types do the programming and own the data. This is what they are trained to do and they do it better than we do it. The folks in IT maintain the data and produce the output summaries but they ought not be forced to try to do that without a statistician's understanding of data.

**Lesson three: Automate and standardize but do so wisely toward a worthwhile end.**

### **So, are all data created equal?**

Yes, all data are created equal in the sense that they all satisfy some particular objective. When stepping outside of that objective, however, caution flags need to be raised. Understanding the effects of this issue cannot be taken lightly. The field of statistics is uniquely qualified to understand the problems that can arise when making this leap. Statisticians need to be involved in all kinds of data analysis and reporting.

### **Where is the future for statisticians in the pharmaceutical industry?**

Statisticians are well-entrenched in the R&D world of clinical trials. The regulatory environment necessitates our existence in that realm. But there are others worlds that could benefit from our training. One such world is Commercial and other areas where there are lots of data but it is being used outside of its primary function. Some of our colleagues already work in that arena but they are vastly underutilized. The opportunity to bring benefit to other areas of the company are ripe for the asking. We can do best for our company and for ourselves by making ourselves available to these other opportunities.