

# An Overview of Data Mining in Drug Development and Marketing

Twenty-sixth Annual Midwest Biopharmaceutical Statistics Workshop  
Muncie, Indiana  
21 May 2003

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

## How did I get here?

In my junior year of college at the University of Dayton, I took my first statistics course. Since it was needed for my mathematics degree, I needed to take it during the semester it was offered, lest I fall a year behind. The problem was that it was only offered during the same time slot as Cell Biology, another course I needed. A biology major coed made easy my choice of which course to attend, so I attended MTH 411 on only four occasions: the tests. The semester proved less than stellar, both with regard to the ladies and the statistics and I decided that statistics just wasn't my bag. The following summer, however, I got a job as a FORTRAN programmer at the University of Dayton Research Institute on campus. I wrote code to do statistical analysis of experiments designed to look for microscopic cracks in jet aircraft engine turbine blades. The statisticians with whom I worked convinced me to try the rest of the statistics sequence my senior year and then to pursue a graduate degree. I headed west to Colorado State University.

Five years later, with a wife, a soon-to-be-born son, and a Ph.D., I headed off to my first job. I had been hired by Marion Merrill Dow, a pharmaceutical company in Kansas City. After nearly four years doing clinical trials in all the preapproval phases of drug development, I moved east to North Carolina and GlaxoWellcome.

At GW (now GSK), I joined a newly formed group called Medical Affairs Medical Data Sciences (MAMDS). The clinical statistics group was split into groups that covered Phase I trials, Phase II–III trials, and Phase IV, or postapproval, trials. MAMDS was this Phase IV group. 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".

All of us who have been doing biostatistics in big pharma for any number of years have served on little project committees at one time or another. Around 1999 or 2000 someone came to the conclusion that GW was spending oodles of money on clinical trials, using the data generated by these trials essentially one time only (for the submission), and then putting the data out to pasture. Surely those data had more value to offer the company, right? Couldn't we leverage this knowledge base somehow? Shouldn't we be able to do something productive with all the data in these databases?

Out of this questioning came MILK, or the MAMDS Initiative to Leverage Knowledge. Since I was known back in those days of my youth as being sort of a maverick (My, how the times change!), I was placed on this committee. Our charter was to recommend strategies for making better use of GW's data assets, to come up with ideas to reuse and recycle the data, to leverage the knowledge in the data.

Around this same time, data mining was starting to become a catch phrase that was being thrown around everywhere you turned. As such, MILK took on a definite data mining direction, at least with regard to trying to document rumors and truths about what data mining was and what it was not. This was my first experience to what is now supposed to be the primary focus of my current job.

A reorganization of the Information Technologies group that supports US Pharma, GSK's US commercial component, led to the development of a group dedicated to doing data mining of the commercial data that GSK purchased or generated. In early 2001, looking for a new environment, the new Data Mining Technologies department looked like something to examine more closely. I officially joined DMT in Summer 2001; we now have four people in our cozy little group. Half of them are statisticians.

### **Is data mining really something different or that just hooey?**

Since joining the data mining group, in addition to attending the traditional statistics meetings, I have also been attending some of the data mining meetings.

A funny thing happens at lots of these data mining meetings. There are always talks and presentations called "What is Data Mining?" or "Are Decision Trees a Statistical Tool or a Data Mining Tool?" or "Why Regression is Data Mining". I, and some others, find this a little odd.

Perhaps these talks are the result of data mining being a relatively young discipline. Perhaps there is a certain amount of legitimacy-seeking on the part of the data mining community. Perhaps no one really knows what data mining is. Perhaps it is nothing more than the confluence of computer science and machine learning and statistics and linguistics and all those other fields.

I think I believe that some time in the not-so-distant past, a computer scientist thought up a funky way to do nonlinear regression. He or she created some seemingly arbitrary functions of the independent variables and then used some additional functions of those new functions to model some outcome. Since the diagram used to describe the algorithm looked a lot like neurons connected to a zillion other neurons, the scientist called this a neural network and said to a statistician, "Hey, Statman, look what I made."

Instead of saying, "Wow, that's interesting; let's check it out and see how we can work together", the statistician said something condescending like "Oh, that's just nonlinear regression. We've been doing that for years. But you're using the wrong nomenclature" and walked away. When other computer scientists saw that this new fancy neural network thingy allowed them to fit models without some snooty statistician getting in the way, they all started doing research on it. Similar things happened with other data mining methods. Thus, the neural nets and decision trees and support vector machines and all the other data mining techniques grew outside of the statistics world. And now that they are getting press in making conclusions based on data, the statisticians want to be involved. Go figure.

There is a data mining meeting this coming Fall called "M2003". Last year there was one called "M2002". Before that were "M2001" and "M2000". An open question to ponder: Will there be an "M2050"? (Will there be a 50<sup>th</sup> annual MBSW?) If the answer is "No", when will the last M meeting be? What will be the writing on the wall when the end is near?

### *Primary and secondary uses of data*

The work that I find myself doing these days isn't a lot different from the work that I did in the past, with a number of exceptions. First, my data mining use of the data isn't the *primary* use of the data. Let me explain.

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. The data make up the transactional record.

Reconciliation of the money in the till is the *primary* use for those data, documentation of a transaction. Their primary use is to document a transaction; that is why they were collected.

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 your merry way. 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 (inventory) 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 but *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.

Both the inventory and customer habits analyses 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. Data are only collected once I go to the pharmacy and turn in my prescription and my insurance information. The data document my transaction. But its secondary uses drive nearly all the statistical work that takes place in the commercial side of big pharma since these data influence much of how sales and marketing do what they do.

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.

Most (but not all) data mining applications are secondary uses of data. I like to call it data recycling. This methodology makes things more prone to issues like bias than in the controlled (clinical) trial setting but that's just the way it is.

#### *Big data sets*

Another factor that seems to segregate the work we do in data mining from the work that the other groups do is the size of the data sets that are used.

Data miners often pride themselves on working with large data sets. When I worked in clinical trials, three-thousand patients was considered a big trial. All those programming tricks concerning efficiency and speed that we learned in SAS classes seemed moot. If reading in *all* the observations takes only 0.072 seconds, who is going to worry, prior to a merge, about reading in only the observations that aren't missing relevant variables. But when you are reading in 147,497,231 observations, it pays to write the code to exclude useless observations before doing merges.

The people working in areas like high-throughput screening also work with very large data sets. A database may hold millions of chemical compounds. Hundreds of variables describe each compound: binding energies, reactivity indices, enzymatic properties, substrate locations, and the like. Doing efficient searching of these types of databases is often considered a data mining problem due to the fact, I think, that one rummages through a pile of data ore looking for nuggets of information. The data were probably not collected to answer a specific question along the lines of a clinical trial; they were collected to fill a database with everything we know about a certain number of compounds.

#### *Disparate data sources*

Another aspect that seems to distance the work I do now from the work I used to do in biostatistics support of clinical trials is the use of auxillary data.

In running a clinical trial, the data needed to answer the questions proposed at the start of the trial are all collected over the course of the trial. If you need it, you had better collect it. If you didn't collect it, you are out of luck.

A question we are working with now at GSK centers around the doctors who call our company's Customer Response Center (CRC) and ask for information concerning who the field sales representative is who supports them.

If you a physician and you prescribe medications for illnesses for which GSK makes remedies, then you will have one or more field sales representatives assigned to you. If you are a really good prescriber of our medications, you will be visited fairly often. If you are not a good prescriber, then you will be visited less often, perhaps never.

When the field sales representatives call on the prescribers, they typically are trying to get a message to the prescribers, something like "Prescribe more GSK drugs, please" or something like that. Sometimes

they will wait hours just to have the chance for a 30-second interaction with a prescriber. It is sometimes very difficult to get in to see the doctor.

Therefore, since it is so difficult to get to see a doc, if a doc actually calls and says “Who’s my rep?”, we ought to send a rep to that doctor immediately, right?

Well, not exactly. The business people know that many docs who call and want to get ahold of a rep are just looking for freebies, like free samples of GSK meds or GSK t-shirts or GSK golf balls. Some of the prescribers don’t write enough prescriptions to make it worth the representative’s time, at least that is the theory.

We’re out to check on the theory. In order to do so, we need lots of different data from different places in the company. We need the CRC call records to find out who called asking about the sales reps. We need prescribing data that tells when the calling docs wrote prescriptions for our drugs and for our competitors. We need data that documents what our reps have been doing and who they have been seeing. Of course, these data are all stored in different databases on different computers owned by different divisions within the company. But to answer the question at hand, we have to pull all that disparate data together.

Many of our data mining investigations seem to have this element of reconstituting data from disparate sources.

*So, is it different?*

No, I think that it is still essentially statistics and here’s why: all the data mining problems I’ve seen have all the essential statistical issues that more formal statistical problems have. They are victims of biases: selection bias, reporting bias, etc. They need to account for variability, be it from sampling or random noise. They need to be cognizant that estimates have errors associated with them and these need to be addressed.

Like all good statistics problems, data mining problems need to be grounded in an eventual action. It is not enough to say “We know that this list of prescribers is going to prescribe more of our drugs if we give them pens and sticky notes.” We need to be committed to following through and taking action. We wouldn’t run a clinical trial and then ignore the results. We ought not do the same with our data mining results.

### **One thing about data wrangling**

One issue that looms large in data mining that may or may not prove so costly in other areas is that of data wrangling. Since our data is typically not collected for the purposes for which we are using it, it often has to go through a data wrangling process to get it into a format that can be analyzed. Typical estimates floating around the data mining community estimate that 70% of a project’s time is used up wrangling the data into shape. I’ll say here that that estimate is probably way, way, way too low. I’ll put the estimate closer to 90%.

The same is true of clinical trials, if you think of the data collection activities as wrangling. If you do, you will see that getting the data ready for analysis (and this includes data generation, collection, scrubbing, and data set preparation), data wrangling is at least 90% of any project’s analysis time.

### **What to take home**

Data mining in drug development and marketing covers a wide array of topics, from evaluating compounds to genetics to study pooling to prescriber targeting. It typically uses data that were collected for some other purpose but it is essentially not different from more formal statistical analyses — it can fall victim to problems with bias and untoward variability and it most definitely suffers all the perils of data wrangling that other investigations do, if not more.