PharmaSUG2011 – Paper MA05

Helping Students Become Effective Industry Statisticians: Supplementing Science with Data Savvy

Aleksandra Stein, Celerion, Lincoln, Nebraska
Steven Kirby, Celerion, Lincoln, Nebraska

Abstract

In the strictly regulated environment of the pharmaceutical industry, a Contract Research Organization (CRO) provides a wide range of services including, but not limited to, “design of a protocol, selection or monitoring of investigations, evaluation of reports, and preparation of materials to be submitted to the Food and Drug Administration.” [FDA 21 CFR 312.3(b)] With regulatory agencies’ recent endorsement of CDISC data submission standards, statistical programmers and statistical scientists now need a solid, functional understanding of data to be effective. Academia, however, has not yet embraced a data-focused training of statisticians. This paper examines the emerging discrepancy between the scientific knowledge students are equipped with and the focus on data required to work in the industry. We also share our methods for enhancing communication and easing the transition between academia and industry.

Introduction

In response to regulatory agencies’ recent endorsement of CDISC data submission standards, the pharmaceutical industry is permeated with requests for data and documentation packages that conform to CDISC SDTM and ADaM standards. Other industries are also increasingly formalizing data requirements. The presence of global data standards is relatively new, but the need for data that are easy to understand and utilize—whether referred to as “clean, analyzable data,” “user-friendly data,” or “robust data,”—is nothing new. With the strong trend towards formal data requirements, it is reasonable to investigate the background and skills needed to thrive in an environment where the science needs to be transparently supported by data that can be traced back to source collection records. The requisite skill set of statisticians in the industry is quickly expanding to include data savvy. Production of data (and associated metadata) which conform to an appropriate structure and meet analysis needs requires a functional understanding of data and effective communication between statistical scientists and programmers. Statistical instruction must impart—at a minimum—a working knowledge of analysis data, the capacity to choose an appropriate dataset structure, and a solid understanding of how to map data into a useful structure without affecting data validity. Academia has remained relative stable in training statisticians to think carefully about analysis without thought to how the data can best support it. In particular, academic institutions have not yet embraced a data-focused educational model.

This paper examines the emerging discrepancy between the scientific knowledge students are equipped with and the focus on data required to work in the industry. We will begin with the diverging definitions/processes of “performing analyses” for statistical programmers and statistical scientists within the industry. Leveraging these characterizations, we illustrate the successes of scientifically oriented higher education as well as the distress to industry caused by the dismal degree to which discussing data occurs in the typical university setting. Specifically, we will discuss the Good (design and analysis), the Bad (data detective work), and the Ugly (dealing with data). Finally, we will mention our methods in enhancing communication and easing the transition between academia and industry.

Analyzing Analysis

Many CRO websites tout their bioanalytical services in terms of design and analysis.

- “designing scientifically-sound and efficient clinical studies and analyzing and interpreting data from these studies”—Celerion
- “adaptive trial design, program and protocol development”—i3
- “Protocol development”, “analysis plans”, and “preparation of statistical reports”—Pharmastat
- “Optimal clinical trial study designs to meet regulatory and promotional needs,” “Statistical analysis support,” and “Analysis plan development”—Ockham

On the other hand, CROs also advertize rapidly available, analysis-ready data.

- “deliver services that enable clients to get products to market faster” and “assist clients in making informed go/no-go decisions on compounds in development as early as possible”—Celerion
- “faster, real-time access to clean clinical data” and “shorter development times and earlier visibility to clean, analyzable data”—Medidata

Statistical scientists address the first set of analytical services: planning, performing, and presenting. Their roles allow them to focus on the big picture—the analysis needed to answer a certain question and the substance of the data that needs to be available to support that analysis. Biostatisticians, therefore, may define analysis as the statistical elements that happen prior to data collection or after obtaining clean, analyzable data.

Statistical programmers, on the other hand, consist of a newer breed of statisticians who take primary responsibility for data preparation and report formatting, facilitating the statistical process. Programmers tend to focus on the details: getting the data prepared for analysis and appropriately formatting data and statistical outputs. Therefore, they perceive analysis as consisting of two parts—preparing the data for analysis and the analysis as defined by the statistical scientists.

By current industry standards, performing statistical analyses would be impossible without both parties working with the same expectations for their data. Examining the time spent on projects post-database lock, one finds that these overlapping tasks (data preparation and analysis) encompass the bulk of billable hours—the majority of these hours being spent on programming in roughly a three-to-one ratio. Industry timelines, therefore, support the interpretation that data manipulation is essential to analysis. Confusion about analysis needs at the data level require a biostatistician's clarification and extend programming, ultimately delaying timelines. Avoiding such delays requires data savvy scientists communicating with their programming support.

Based on the core curriculum offered by statistics departments at universities nation-wide, however, we are forced to conclude that academia supports the constricted portrayal of statistical analysis limited to textbook-quality data. Given these incongruous definitions, recent students occasionally find themselves underprepared to function efficiently in an industry which requires statisticians to transform one set of data into safety reports, efficacy reports, integrated data, submission compliant data, analysis data, and more! We therefore need to take a closer look at what higher education is doing right and where it could use a little help in helping students.

Data Too Good to Be True

Opening ten stats textbooks in the nearby university library, I found the same information on randomization, regression, power, and error rates as in the 2007 *Pharmaceutical Statistics Using SAS®: A Practical Guide*. Granted, each work differed in organization and presentation, but all contained the appropriate information needed for understanding statistical designs and analyses, as mentioned above.

The Good: Analysis and Reporting

While adjusting to a company's methods, software, and related specifications may be tricky, basic design and analysis are stable across studies and between companies. This is hardly surprising; ANOVAS and T-test procedures are invariant to data from lions, tiger, bears, and chemical compounds. Similarly, finding confidence intervals depends on underlying distributions, sample sizes, and desired levels of confidence, not on the subjects from which the data are taken. Moreover, parallel studies and crossover designs apply equally well to fruit flies, racehorses, and Alzheimer patients. Statistical design and analysis applications may vary by discipline but the tools taught in the classroom and classic textbooks are consistent. Whether they studied statistics from a department emphasizing animal breeding and agriculture or arthritis and abdominal cancer, students absorb appropriate design and analytical methodologies in academia.

In terms of teaching students to communicate, schools score favorably in this aspect as well. Academic practice in explaining experimental designs, sampling schema, and interpretations adequately prepares incoming industry statisticians to customize company SAPs and protocols. Cultural and language barriers notwithstanding, university coursework emphasizes the importance of reaching and reporting research conclusions. The transition from freeform interpretations to standard tables, listings, and graphs nicely tied into report templates is generally graceful. While verbal communication skills will vary, including the ability to talk with co-workers and clients, statistical scientists should be ready to write.

Based solely on the design, analysis, and reporting aspects which CRO websites advertize, academic institutions produce students who may—with guidance and industry-specific enlightenment—fulfill the roles of pharmaceutical biostatisticians. Most CRO websites, however, fail to mention that sometimes someone somewhere might make a mistake. In particular, sometimes mistakes find their way into a statistical scientist's dataset—a phenomenon that rarely occurs in statistical students' data.

The Bad: Data Detectives

A closer look at the statistical library contents reveals that textbooks mainly contain textbook data: complete, balanced data presented neatly in rows and columns ready for analysis. Class assignments tell the same tale of perfect, simulated data for homework. Unlike many academic disciplines, the pharmaceutical industry is strictly regulated by the FDA who expects data to be 100% accurate. Missing or unusual observations must be adequately explained—and academia is not relating this message. Of course, universities and books mention here and there that a statistician may wish to remove an outlier from a dataset or work with the median rather than the mean to minimize its effects. In the pharmaceutical industry, however, neither of these responses would be appropriate in most situations. Instead, a biostatistician (programmer/data manager) would flag the dataset or patient information containing inexplicable or surprising information and do some detective work. Prior to FDA submission, said statistician would need to discover whether or not the information was correct, then either track down accurate information for reporting or explain the results. Pharmaceutical veterans perform this process routinely, but beginning biostatisticians are often ill-equipped for data-driven error detection and explanation. Outliers, unbalanced data, and missing values are the tip of the iceberg for dealing with data. Even clean data which are not analysis-ready may stump inexperienced statisticians.

The Ugly: Dealing with Data

Except in special cases with special relationships, university statistics departments offer courses in statistics. Their curricula rarely cover courses in data beyond data-mining; as mentioned above, the data students do see are nearly out-of-this-world perfect. We propose here a list of the top three astounding areas to which recent graduates did not receive enough exposure.

1. ***Structuring Data.*** Analysis data don't happen by accident. Clinical data, for example, is most efficiently collected using a horizontal structure, below.

Patient	Systolic 1	Diastolic 1	Systolic 2	Diastolic 2	Systolic 3	Diastolic 3
001	118	76	117	76	110	71
002	142	96	156	101	145	83
003	135	80	137	82	129	76

Horizontal Data Structure

As per CDISC guidance for CDASH, SDTM, and ADaM, clinical data may be most effectively analyzed, reported, and warehoused using a vertical structure—and most often not a direct transposition.

Patient	Test	Result
001	Systolic 1	118
002	Systolic 1	142
003	Systolic 1	135
001	Systolic 2	117
002	Systolic 2	156
003	Systolic 2	137
001	Systolic 3	110
002	Systolic 3	145
003	Systolic 3	129
001	Diastolic 1	76
002	Diastolic 1	96
003	Diastolic 1	80
001	Diastolic 2	76
002	Diastolic 2	101
003	Diastolic 2	82
001	Diastolic 3	71
002	Diastolic 3	83
003	Diastolic 3	76

Vertical Data Structure

Transformation tools such as PROC TRANSPOSE or PROC SQL, however, are not needed for analysis of classroom data and thus are not introduced in the classroom.

Getting data into and out of a multivariate data structure is another incredibly common but overlooked obstacle in academic data structuring.

2. ***Sorting Data.*** Statistics student rarely explore data beyond summary statistics. Therefore, even if they have heard of PROC SORT, a student will not be familiar with nuances of sorting with respect to numeric or character variables. A thorough tutorial was given last year by Andrew Kuligowski. For larger datasets, a simple sort will be insufficient and sorting must be done BY variables, perhaps with enhanced functionality

such as NODUPKEY sorts. Of course, there are also occasions when a subset of a dataset is desired and WHERE or IF statements are necessary.

Similar issues arise for ranking data, say, for nonparametric analysis: educational examples focus on analysis and are small enough to rank by hand. PROC RANK would eliminate manual imputations and their associated errors, especially under the stress of impending deadlines.

2.5 ***Merging Data.*** A sister issue to sorting is merging—which merits mention, but not a new number. When each particular problem that appears in a semester-long course is accompanied by a specific simulated dataset, merging is never necessary.

Merging datasets together or merging in specific variables requires indentifying key variables or specifying that data are unique at the merge level. Using CDISC terminology, a statistician must know whether the datasets in question contain one entry per subject, one entry per event per subject, one entry per event per visit, etc. David Franklin's 2010 Tutorial enumerated popular merging methods ranging from a simple datastep MERGE with a BY option to merging with PEEK and POKE functions.

3. ***Tracking Data***. In the pharmaceutical industry, the FDA requires completely documented data. Following this trend, unregulated industries and academia are now knee deep in discussions of reproducible data. Outside of SAS—where leveraging the built-in functionality of data labels is trivial—labeling, formatting, and commenting on data (or, rather, lack thereof) has been resulting in irreproducible data. Adding a variable LABEL; coding with PROC FORMAT or IF/THEN and ELSE statements; and documenting changes to data (sorting, categorizing, transforming, etc.) in comments or in the program header relieve much worry in terms of creating reproducible research, in the pharmaceutical industry and in other industries soon to follow.

If you strip a statistical programmer of these core data concepts or ban a biostatistician from using these tools to describe and create analysis-ready data, no CRO would advertize "real-time client access" to data. In addition, no CRO would be capable of keeping up with demands for rapid decision-making and drug-development data. Thinking about data, talking about data, and dealing with data are essential elements of statistical analysis in the pharmaceutical world.

C-ROck: A Case Study

C-ROck is a young company with an old history and a nearby university, UNI. For decades, many of C-ROck's statistical employees have originally been students in the Department of Statistics at UNI. Unfortunately, C-ROck has been experiencing concerns with the growing discrepancy between its needs and the skills of UNI graduates. Coincidentally, several scientists at C-ROck decided to audit a statistical course at UNI as a refresher. This helped the statistical team at C-ROck to identify particular areas of concern regarding statistical education at UNI with respect to pharmaceutical industry skills and simultaneously strengthen personal relationships between professors and professionals at the two locations.

Subsequently, the instructor of the "refresher course" decided to teach a Pharmaceutical Statistics class and moreover invited several members of the C-ROck to be guest speakers during the course. Both parties benefited from this semester: the students gained a real-world perspective and appreciation for the importance of data in the statistical world which are underemphasized in academia; C-ROck eventually gained new employees armed with statistical science and data savvy. Additionally, previous C-ROck statistical programmers and the new hires collaborated in an increasingly rigorously regulated environment, inspiring innovative standard processes to efficiently process submission packages for CDISC SDTM and ADaM with Define.XML capabilities. Finally, persons from both sides continued to enjoy perks of the partnership such as continuing education for biostatisticians who bring to the classroom a perspective of industry experience; availability of part-time and temporary statistical employees who fill a need and receive funding and experience in return; and, of course, extended networking opportunities and success stories.

Conclusion

Whether in SAS or any other software, a statistical student would be hard-pressed to receive data-as-collected and transmit data-as-analyzed without leveraging the above resources for dealing with data. Nevertheless, academia has yet to embrace a data-driven statistical education model, resulting in repeated rough transitions into industries across disciplines. While we would admire attempts to revolutionize higher education, we are not—at this point—advocating it. Instead, we have presented our approach to smoothing the passage from student to professional, on the pharmaceutical end.

Being aware of the data-driven nature of statistical analysis in industry and armed with the knowledge that analysis is the sole focus of academia, new hires weaknesses are now anticipated. More effective training can be planned and teams structured to alleviate this situation, maximizing the efficiency of new hires and speeding their transition into data savvy scientists.

Acknowledgments

My education would not be complete and none of this would be possible without the cooperation and support of Celerion, Inc., and the Department of Statistics at University of Nebraska-Lincoln. Thanks!

Recommended Reading

- Dmitrienko A, Chuang-Stein C, D'Agostino R. *Pharmaceutical Statistics Using SAS®: A Practical Guide*. Cary, NC: SAS Institute, Inc. 2007.

- Cody R P, Smith J K. *Applied Statistics and the SAS® Programming Language.* Fifth Ed. Upper Saddle River, NJ: Pearson Prentice Hall. 2006.

- Kuligowski, A. TU07: *The Ins and Outs Ups and Downs of Sorted Data*. PharmaSUG 2010.

- Franklin, D. TU06: *Countdown of the Top 10 Ways to Merge Data*. PharmaSUG 2010.

Contact Information

Feedback is welcome. Please contact the author at Aleksandra.Stein@gmail.com.