Data Sharing and the Development of the Cleveland Clinic Statistical Education Dataset Repository

Amy S. Nowacki
Cleveland Clinic

Journal of Statistics Education Volume 21, Number 1 (2013),
www.amstat.org/publications/jse/v21n1/nowacki.pdf

Key Words: Data Sharing; Teaching Examples; Educational Resource; Biostatistics; Medical Data; Clinical Trial Data; Health Data.

Abstract

Examples are highly sought by both students and teachers. This is particularly true as many statistical instructors aim to engage their students and increase active participation. While simulated datasets are functional, they lack real perspective and the intricacies of actual data. In order to obtain real datasets, the principal investigator of a study must be willing to share the data. Understanding investigators' opinions regarding data sharing would thus help elucidate the general lack of data sharing currently exhibited. Presented are the results of a survey designed to gather information regarding the proportion of researchers willing to share their data, conditions, formats, primary motivation, concerns and current availability of data for sharing. With 76% (56/74) responding favorably to the idea of sharing their published data, the creation of a new statistical educational resource was prompted. Thus, additionally described is a web-based dataset repository that can be used as a resource by both educators and students of statistics. This growing repository presents raw data from real medical studies and offers (a) a vignette summarizing the study, research question and study design; (b) a data dictionary with clear documentation of variables and codes; (c) a complete citation for the associated study publication; and (d) a variety of data formats compatible with the majority of statistical packages. The repository went online on 12/18/12 at the URL http://www.lerner.ccf.org/qhs/datasets/.

1. Introduction

Many trained biostatisticians perform multiple roles. A common scenario is to be both a collaborative medical researcher and a statistical educator. As a collaborator, time is spent actively involved in study design, data collection, data preparation, data analysis, and reporting

of medical studies. In this environment, we work daily with real research questions and real data. Hence, datasets are abundant and plentiful. As an educator, I often spend an inordinate amount of time searching and preparing datasets for use in class. I found this to be particularly true after a recent transition to a problem-based learning approach to teaching Introductory Biostatistics (Nowacki 2011). Within these dual roles appear both the problem and the solution.
According to Gladwell (2005), "We learn by example and by direct experience because there are real limits to the adequacy of verbal instruction" (p. 70). Examples are highly sought by both students and teachers as their role in learning is critical (Chi, Bassok, Lewis, Reimann & Glaser 1989; Quilici & Mayer 1996). The benefits of problem-based learning have been established (Weimer 2007; Visconti 2010), but an often overlooked aspect of implementation (particularly for statistics) is the availability of data. While simulated datasets are functional, they lack real perspective and the intricacies of actual data. In addition, the use of real data is in alignment with the constructivist learning environment that emphasizes the complexity of the real world and having students perform authentic tasks in a meaningful context (Bransford 2000; Mvududu 2005). Arguments have been made for using real data so that students can experience the delight of answering a real-life research question and start to understand why certain methods are used, not just how they are used (Willett & Singer 1992). The use of real data is also a recommendation of the Guidelines for Assessment and Instruction in Statistics Education (GAISE) report (American Statistical Association 2012). Specifically, one recommendation for teachers is to search for good, raw data to use from web data repositories, etc.. The report asserts that more emphasis on data can improve any statistics course and that classes should rely more on projects, lab exercises, and group problem-solving and discussion activities, all of which require data. In order to obtain real datasets, the principal investigator of a study must be willing to share the data. Understanding investigators' opinions regarding data sharing would thus help elucidate the general lack of data sharing currently exhibited.

This article describes the development of a new resource for statistical educators providing clean, de-identified and well-annotated medical datasets. In the process of assessing the feasibility of such a resource, we conducted a survey study aimed to estimate the proportion of investigators willing to share data as well as characterize what influences this decision.

2. Methods

2.1 The Website

A number of websites currently make datasets available to others. Notably, there is the Vanderbilt Department of Biostatistics Dataset site (Harrell 2011), Carnegie Mellon's Data and Story Library (DASL 1996), and the Chance database (Snell 1999), and each is effective for serving its own individual purpose. So why create a new resource? Attributes that enhance the instructional suitability of a dataset have been described previously (Willett & Singer 1992). Notably mentioned are datasets that come in raw form, are authentic, include background information, are intrinsically interesting or relevant, are topical or controversial, and lend themselves to various statistical analyses. Thus, it is desirable to have a dataset repository that offers raw data from real medical studies providing a study vignette, clear documentation, and several variables.

2.2 The Survey

A brief survey was created by the author to gather information regarding the proportion of researchers willing to share their data, conditions, formats, primary motivation, concerns and current availability of data for sharing (see the Appendix for a copy of the survey instrument). Survey questions were reviewed by a questionnaire design expert to ensure they adequately addressed the research questions. Potential participants received an email invitation to complete the two minute survey. Names of those invited were obtained from a master list of all past or current research preceptors for students at the Cleveland Clinic Lerner College of Medicine (CCLCM). Research preceptors are MD or PhD investigators who have a project idea and agree to mentor a medical student for either a summer (approximately 9 weeks) or during the student's research year. Thus these are individuals with familiarity and involvement in medical research and most likely to possess datasets of interest. The survey introduction explained that the goal is to build a repository of data (open website) from Cleveland Clinic research studies that will: (1) act as a resource for our instructors and other statistical educators; (2) showcase the basic, clinical and translational medical research of this institution; (3) foster future collaborative efforts, and (4) protect their intellectual property while allowing others to learn from their work. The survey invitation was sent to approximately 175 individuals with 74 (42%) participating. Of those completing the survey, 80% (59/74) were CCLCM faculty members. Results of the survey were captured in a REDCap secure database. This survey was conducted with Cleveland Clinic IRB approval.

3. Results

Initially, each participant was asked if he or she would ever consider granting permission for his or her de-identified data to be posted to a Cleveland Clinic medical study statistical repository (open website) for which 76% (56/74) responded favorably to this question. This was extremely encouraging and enough of a response to deem the resource creation viable.

The survey next investigated, among those responding favorably to data sharing, what conditions they would require. Sixty-four percent (36/56) would necessitate that the study already be published. This later became a requirement of contributions to the repository in an effort to protect the intellectual property of our investigators. Ninety-six percent (54/56) would require a disclaimer be agreed to by end-users before download of data. Below is the current disclaimer and policy for usage:

> "I understand that the datasets of this educational resource have been de-identified and I agree not to claim or imply that any inferences (beyond those of the study's original purpose) derived from these educational datasets are valid estimates."
>
> "Datasets may be freely used in teaching without contacting the author. Datasets should be cited giving the name of the dataset, the lead author, and the date accessed. To use datasets in papers, books, or other published material you must obtain the consent of the lead author."

Seventy-seven percent (43/56) would require that a registration be completed by end-users before download of data. This was implemented with a very brief 5 question registration that allows us to collect information regarding resource usage and incorporate a disclaimer agreement by end-users. The goal is to be very simple and noninvasive as we did not want this to become a deterrent to usage. Instructors may register and download data making it available to only their students online if the site is secure, and this will avoid each student having to register individually. Participants were also given an opportunity to submit additional conditions, but none notable were recorded.

Shared data tends to fall into one of two categories: complete or random subset only. Those who share complete datasets often note the importance of reproducibility in science and that reproducing published results is a powerful learning experience. Those who share random subsets only often allude to the protection that this offers and that while one cannot achieve exact replication of published findings, they can learn the process and verify overall conclusions. The majority of survey participants, 71% (40/56), indicated that they would prefer the format of submitting a complete de-identified dataset.

When asked about their primary motivation for agreeing to share data, 39% of respondents (22/56) selected contribution to scientific community, 23% (13/56) selected that the study data will provide extraordinary learning experiences for students, 20% (11/56) selected to emphasize that research is done for the benefit of our patients, not ourselves, 14% (8/56) selected the possibility of leading to new techniques or findings through collaboration, 1 individual (2%) selected additional exposure of their research, and 1 individual (2%) selected other and wrote "To help the medical school."

The survey then inquired about concerns regarding data sharing. Responses among those in favor of sharing data covered a broad spectrum of opinions. Thirty-six percent (20/56) noted no concerns. One participant wrote the following, "I have no concerns, you can't publish what is already published." Alternatively, 57% (32/56) were concerned that others might publish something that would misrepresent their findings, 23% (13/56) were concerned that they would not properly be credited for the data, and 1 individual selected Other and wrote "Time involved in pulling together a dataset for teaching purposes is not extra time that I have." If a participant indicated that he or she would not be willing to share their data, this was the only other question asked of him or her. Among this subset, the most cited reason 50% of participants (9/18) gave for not considering sharing data was concern about protecting intellectual property. Each of the following were also reasons provided for not sharing data: lack of time, concerns with Internal Review Board (IRB), concerns with sponsorship, misinterpretation of data by others, and an inability to de-identify data in their field.

Finally, when those in favor of sharing data were asked whether or not they currently possessed data of a published study, 27% (15/56) responded yes with 14 providing email addresses to be contacted further. Meetings were held with each of these investigators providing an opportunity to explain in detail what a dataset contribution would entail and the types of datasets appropriate for the repository. As a result, some datasets were deemed inappropriate (e.g. utilizing data bought by the investigator, study not yet published, etc.) while others are in preparation for upload.

The process is slow, but the resource is growing and will be a continued work in progress. The link to the resource is as follows: http://www.lerner.ccf.org/qhs/datasets/. The website provides study vignettes providing brief descriptions of the research question, the study design and study sample. A data dictionary is also available providing clear descriptions of variables including units of measure and categorical coding. Original manuscript citations are provided so that users can access more study detail or for instructors who wish to have students replicate study findings. A link to the user registration is required before data download and once completed the user is given a new URL taking them to a website where all datasets are available. The data are offered in three formats for easy import into most statistical packages (*.xlsx, *.sas7bdat, *.csv).

4. Discussion

Utilization for illustrative demonstrations during class, homework assignments, and student projects or exams often exhausts the limited supply of user-ready real datasets openly available. The idea of a dataset shortage is often met with disbelief as challengers site advances such as mandatory clinical trial registration. It is important, however, that a distinction be made between making study results available and making study data available, the latter remaining quite rare (Chan 2011). While it is true that other initiatives such as the National Institutes of Health policy on data sharing (National Institutes of Health Policy on Data Sharing 2003) and the efforts of some journals to require authors to share their data has helped improve data availability, these policies remain largely unenforced. A recent study sought to determine how well authors comply with such journal policies by requesting data from authors who had published in journals with clear data sharing policies (Savage & Vickers 2009). The result was that only one of ten raw datasets requested were received, suggesting that such data sharing policies do not lead to authors making their datasets available to others. Clearly, obtaining access to original datasets remains a challenge.

In acknowledging that there are real and perceived obstructions to sharing raw data (e.g., patient privacy, authorship and future publishing opportunities, uncovering an error, employing alternative analytic methods, etc.), some simple guidelines exist to protect the rights of investigators (Smith 1994; Kirwan 1997; Vickers 2006). Agreeing with the authors of these guidelines, investigators should be included as co-authors on any publication resulting from the re-analysis of raw data or, alternatively, be offered the opportunity to provide a response and commentary.

An added benefit of depositing study data to this repository is that it would fulfill requirements of open data access for our contributors. Many funding agencies require a description of how the data will be made available and such a resource is in direct alignment with that intention.

5. Conclusion

Data sharing concerns are widespread. Learning what concessions encourage data sharing will benefit all involved in statistical education as well as other areas. Effective data sharing includes communicating that data are available, providing sufficient descriptive information about the data, rendering the data in a usable format, and making data accessible. The Cleveland Clinic Statistical Education Dataset Repository achieves this high standard and it is our hope that others will utilize and benefit from this resource. Future plans include expanding the repository to allow outside dataset submissions with the goal of having a Cleveland Clinic collection alongside a contributed collection. Additionally, there are plans to generate a bank of educational materials associated with the datasets to assist or inspire statistical educators. This repository will continuously be appended to and enhanced.

Appendix

Cleveland Clinic

Willingness to Share Data Survey

1. When sharing data, numerous techniques exist to protect your intellectual property; for example, only sharing published data, only sharing a subset of the data, requiring users to agree to a disclaimer or register before downloading data, etc. With that in mind, would you ever consider granting permission for your de-identified data to be posted (with proper recognition to you and instructions for proper citation) to a Cleveland Clinic medical study repository (open website)?

○ Yes

GO to Question 2

○ No

GO to Question 10

2. As a condition of your agreement to share your de-identified data, would you require that your study manuscript already be accepted for publication?

○ Yes

○ No

3. As a condition of your agreement to share your de-identified data, would you require that a disclaimer (such as the example below) be agreed to by end-users before download of data?

"I understand that the teaching data has been rendered anonymous through the application of certain statistical processes. I agree not to claim or imply that any inferences derived from the teaching datasets are valid estimates. If I intend to use the data for purposes other than teaching, I agree to obtain permission and the original complete dataset from the listed contact author before using in any publication or presentation"

○ Yes

○ No

4. As a condition of your agreement to share your de-identified data, would you require that a registration be completed by end-users before download of data?

○ Yes

○ No

5. Do you have any additional condition(s) that you would require before agreeing to share your de-identified data?

○ Yes

○ No

6. Which format would you prefer to provide your de-identified data in?

○ complete de-identified dataset
- reproducibility in science is important
- students learning how to reproduce published results is a powerful learning exercise

○ random subset of de-identified dataset only
- this option affords more protection
- while students can not exactly replicate my results, they will have enough data to learn the process and verify my conclusions

7. What is your primary motivation for agreeing to share your de-identified data (select one)?

○ to emphasize that research is done to benefit our patients, not ourselves

○ contribution to scientific community

○ the study data will provide extraordinary learning experiences for students

○ additional exposure of my research

○ sharing of data may lead to techniques or findings or further research collaborations

○ other ____

8. Do any of the following concern you regarding sharing your de-identified data? (check all that apply)

□ no concerns

□ concerned that others might publish something that would misrepresent my findings

□ concerned that I will not be properly credited for the data

□ other ____

9. Do you currently have the data of a published study that you would be willing to share?

○ No I have not yet published a study appropriate for sharing data

○ Yes I published a study and would be willing to share the de-identified data

STOP

10. Since you elected not to share your de-identified data (in any capacity), what is your main reason (biggest concern) and is there any way that we could address it?

STOP

References

American Statistical Association. GAISE reports. [cited 2012 Dec 1]. Available from: URL: http://www.amstat.org/education/gaise.

Bransford, J. (2000), *How people learn*. Washington, DC: National Academy Press.

Chan, A. (2011). "Access to clinical trial data," *British Medical Journal*, 342, d80.

Chi, M.T.H., Bassok, M., Lewis, M.W., Reimann, P., and Glaser, R. (1989), "Self-Explanations: How Students Study and Use Examples in Learning to Solve Problems," *Cognitive Science*, 13, 145-182.

DASL: Carnegie Mellon University Data and Story Library. (1996), from http://lib.stat.cmu.edu/DASL/.

Gladwell, M. (2005), *Blink: The Power of Thinking Without Thinking*. Boston, MA: Back Bay Books.

Harrell, F. (2011), from Vanderbilt Department of Biostatistics http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets.

Kirwan, J.R. (1997), "Making Original Data from Clinical Studies Available for Alternative Analysis," *The Journal of Rheumatology*, 24(5), 822.

Mvududu, N. (2005), "Constructivism in the Statistics Classroom: From Theory to Practice," Teaching Statistics, 27(2), 49 - 54.

National Institutes of Health Policy on Data Sharing. (2003), from http://grants.nih.gov/grants/policy/data_sharing/.

Nowacki, A.S. (2011), "Using the 4MAT Framework to Design a Problem-based Learning Biostatistics Course," *Journal of Statistics Education*, 19(3), 1 - 24. http://www.amstat.org/publications/jse/v19n3/nowacki.pdf

Quilici, J., and Mayer, R.E. (1996), "Role of Examples in How Students Learn to Categorize Statistics Word Problems," *Journal of Educational Psychology*, 88(1), 144-161.

Savage, C.J., and Vickers, A.J. (2009), "Empirical Study of Data Sharing by Authors Publishing in PLoS Journals," *PLoS ONE*, 4(9), e7078.

Smith, G.D. (1994), "Increasing the Accessibility of Data," *British Medical Journal*, 308, 1519 - 1520.

Snell, J.L. (1999), from Dartmouth College CHANCE Database http://www.dartmouth.edu/~chance/teaching_aids/data.html.

Vickers, A.J. (2006), "Whose Data Set is it Anyway? Sharing Raw Data from Randomized Trials," *Trials*, 7, 15.

Visconti, C.F. (2010), "Problem-Based Learning: Teaching Skills for Evidence-Based Practice," *Perspectives on Issues in Higher Education*, 13(1), 27 - 31.

Weimer, M. (2007), "Problem-based Learning: Benefits and Risks," *The Teaching Professor*, 21(2), 5.

Willett, J.B., and Singer, J.D. (1992), “Providing a statistical ‘model’: Teaching applied statistics using real-world data,” In F. Gordon and S. Gordon (eds.) *Statistics for the Twenty-first Century*. Washington, DC, Mathematical Association of America.

Amy S. Nowacki, PhD
9500 Euclid / JJN3 - 01
Cleveland, OH 44195
Phone: 216-444-3773
Fax: 216-444-8021
Email: nowacka@ccf.org
