The TSHS Resources Portal: A Source of Real and Relevant Data for Teaching Statistics in the Health Sciences

1. BENEFITS OF USING REAL AND RELEVANT DATA

It has been widely acknowledged for many years that using real data in teaching statistics is more effective than using "toy", made-up, or context-free data (Garfield et al. 2005). The American Statistical Association-endorsed Guidelines for Assessment and Instruction in Statistics Education College Report (Everson, Mocko, et al. 2016) contains six recommendations for effective teaching at the undergraduate level. The third recommendation states "Integrate real data with a context and a purpose" and goes on to say that "Using real data in context is crucial in teaching and learning statistics, both to give students experience with analyzing genuine data and to illustrate the usefulness and fascination of our discipline." For teaching in the health sciences, we find that the most engaging, and therefore ideal, data are those which are both real *and relevant* to the health sciences. Better still are datasets that are associated with published research articles. Students engaged with such data learn to read the published article, analyze the data, and make connections between the two.

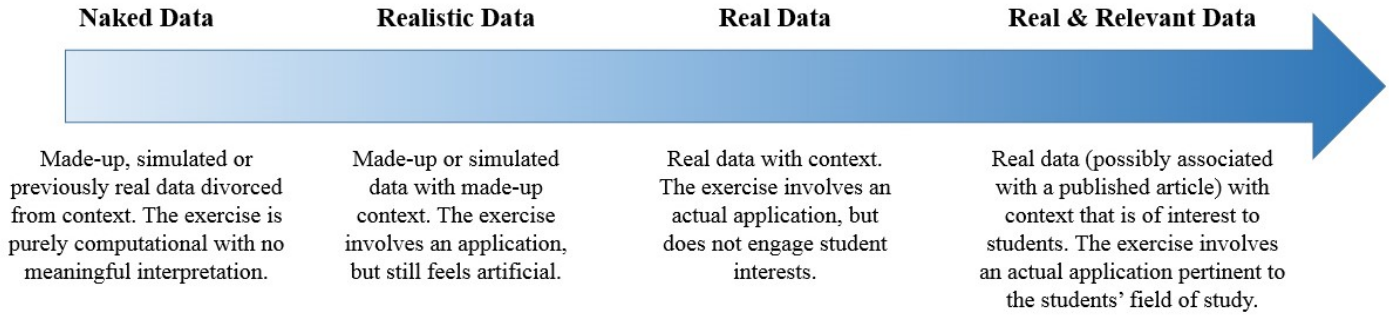

Figure 1. Spectrum of utility for instructional data

The GAISE guidelines (Everson, Mocko, et al. 2016, pages 61-62) describe a spectrum of data "reality" (summarized in Figure 1) that ranges from "naked" data to "real and relevant" data. "Naked" data are datasets that consist only of numbers, with no units and no context. This could include not only made-up, synthetic or simulated data but also data from a real study where the context has been omitted or lost after many years of being passed around from instructor to instructor. "Realistic" data refers to made-up or simulated datasets with a bit more information; relative to "naked" data, these provide information about the putative context, such as units, variable labels and a sentence or so about the application. "Realistic" data may also include so-called "toy" datasets: small, simplified datasets used to illustrate a specific limited point. "Real" data refers to datasets from an actual study with full context provided. A limitation of "real" data is that they may or may not be of direct interest to the student. Finally, "real and relevant" data are datasets which are distinguished for being both from an actual study and of direct interest to the students (that is, pertinent to their field of study).

We advocate for "real and relevant" data for teaching in the health sciences. In particular, we recommend the use of data that were collected in an actual medical or public health-related research study, ideally accompanied by the published study article. If impractically large, such a dataset can be subsetted to achieve a usable size. Also, as needed, such data can be cleaned or partially cleaned prior to its use in teaching, to avoid frustrating novice students with data cleaning tasks. A caveat is that, for human subjects protection reasons, real and relevant data must be de-identified and HIPAA-compliant prior to sharing it with students. Collectively, any manipulations of the source data must preserve the "real and relevant" context and authenticity.

An example of "real and relevant" data is the Licorice Gargle Dataset, which is available on the TSHS Resources Portal (Nowacki 2017). It contains data on 236 adult patients undergoing elective thoracic surgery requiring a double-lumen endotracheal tube. Participants were enrolled in a randomized controlled trial that compared the effectiveness of a licorice gargle with a sugar solution gargle for the control of post-extubation coughing and sore throat. This dataset can be used to illustrate contingency tables, logistic regression modeling, and the more advanced topic of ordinal logistic regression. The published article describing the study is referenced in the Portal as well (Ruetzler et al. 2013). If desired, the Licorice Gargle Dataset could be modified or subsetted to yield teaching datasets that are to the left in the spectrum of utility (Figure 1), but in our view, the Licorice Gargle Dataset is best used as real and relevant data; it gives students who are interested in health sciences the complete dataset, the data dictionary, an introduction to the study (all available on the TSHS Resources Portal), and the published study article.

1.1 THE VALUE OF REAL AND RELEVANT DATA

The value of using real and relevant data is particularly apparent when teaching professionals who will encounter such data in their work. In our experience with teaching in the health sciences, using real and relevant data benefits the learner in multiple ways.

Relevance

The notion of relevance is the idea that at some point in the course of teaching, the students should actually come to regard the problems introduced by the instructor as their own problems (Libman 2010). Real and relevant data resonate with the current or expected experiences of the student and therefore are likely to be especially motivating and interesting (Scheaffer 2001, Garfield & Ben-Zvi 2009, Neumann, Hood, Neumann 2013). The use of real and relevant data is likely to preclude the common student lament, "how will I use in real life what I'm learning in this class?" because the answer is already clear. An appreciation of real world relevance is particularly valuable for students who are not statistics majors; they are moved from experiencing statistics as an abstract field to recognizing statistics as an important skill set related to their job and career expectations.

Furthermore, we believe that when teaching statistics in the health sciences, it is important to communicate that statistics in real world health sciences research exists within a complex, collaborative, and highly structured scientific framework (Enders et al. 2017). Specifically, teachers should facilitate students gaining an appreciation of the fact that high quality research begins with an important scientific question that is developed into a detailed research proposal. Optimally, students will acquire an understanding of the need for detailed and deliberate study planning in order to execute the research proposal and produce the final dataset that will be used for all statistical analyses. Although the full complexities of this scientific framework are beyond the scope of most statistics courses, providing introductions to this scientific context through the use of real and relevant data is invaluable. It not only engages students but also exposes them to the key concept that much research is driven by questions and hypotheses that are developed prior to the collection of data. Such real-world appreciations are not easily conveyed by "toy" or simulated datasets.

Another advantage of using real and relevant data from the students' area of expertise is that it can reduce the students' anxiety level. They may not be experts in statistics, but they may well be experts (compared to the instructor) in another field such as surgery, nursing, pharmacy or hospital administration. On this point, this is an aspect of teaching statistics in the health sciences that distinguishes it from teaching undergraduate statistics. The instructor is an expert in statistics but at best an experienced amateur in the content area (e.g., oncology, public health, dentistry, veterinary medicine). By contrast, the students are typically expert in their own fields but not in statistics. This evens the playing field between instructor and student and makes the relationship much more collegial and collaborative. Use of real and relevant data can thus make it easier for students to contribute to class discussions.

An additional "relevance" to the use of real and relevant datasets is that it trains students to become knowledgeable statistical consumers. Informed statistical consumers see the relevance of statistics to their future on-the-job activities and, quite likely, are increasingly motivated to continue learning statistical concepts. By contrast, when instructors fail to convey the practical importance of statistics, students rarely become invested. In sum, we believe that the easiest method for conveying practical importance is to demonstrate relevance.

Meaningful assignments

Real and relevant data support the generation of assignments that are meaningful (Willett & Singer 1992, Briggs 2014). In our experience, meaningful assignments are preferable to those that are contrived "busy work" or simplified assignments. Instead of having the intended effect of making the concept easier to learn, "busy work" assignments often have the unintended effect of frustrating students because there is little or no perceived connection to the work activities of their intended field. One example of a meaningful assignment approach is the "see one/do one/evaluate one" type of assignment (Samsa, Lee, Neal 2012). To illustrate, consider Table 2 in the Ruetzler et al. (2013) article accompanying the Licorice Gargle dataset. It presents the relative risk of sore throat at three time points and compares licorice gargle patients relative to sugar water control patients. The instructor might lead a walkthrough of the analysis of the first time-point in an illustration of the procedure. The students are then tasked to do the analysis of one or both of the other time points during an in-class data analysis activity. Finally, students can evaluate a similar analysis in a different published article on a similar health topic to aid transference of their knowledge. At the conclusion, the students have hopefully learned, applied, and understood the new method in the context of their field of interest.

Demystification

A substantial benefit of real and relevant data occurs when the study findings have been published. Simultaneous access to both the data and the published findings provides a unique opportunity to practice the steps involved in managing collected data, performing statistical analyses on those data, and finally synthesizing the results of statistical analyses in succinct reports suitable for publication (Smith 2008, Everson, Mocko, et al. 2016). Absent such "practice runs", students are unlikely to possess the skills needed to, first, navigate an overabundance of statistical software output and then, subsequently, produce concise and intuitive tables and figures akin to those found in reports in the literature. Students new to statistics are quite likely to be mystified by the apparent disconnect between the overabundance of output and the brevity of published reports. Thus, we recommend demystifying the publication process. For example, when students are given the Licorice Gargle dataset with its published manuscript, they can reproduce results presented in the paper, allowing them to translate the software output into the tables and figures presented in the paper. This gives them a better understanding of where the results came from.

Reviewing the published work also provides an opportunity to see statistics in the context of research. As an example, Figure 1 of Ruetzler et al. (2013) presents the CONSORT participant flow diagram for the study, facilitating a conversation about trial design, data analysis plans, and proper reporting (Schulz et al. 2010). The study design, testable hypotheses, valid data collection tools, appropriate analyses, transparent presentation of results, and final conclusions are all components of research that involve statistical concepts. Students, so engaged, realize that the authors of the paper did exactly what they themselves are learning to do: they took a dataset, did calculations on it, and summarized the results to obtain the tables/figures in the paper. This kind of assignment can also provide a segue into talking about reproducible research principles.

Data Management

Real data is often messy and undocumented. Anyone who has worked as a statistical consultant or collaborator with non-statisticians is likely aware that very few of their clients understand how much time data management takes or how they, the clients, could structure their datasets in order to minimize this time (Broman & Woo 2017). This makes instruction in the management of real data a valuable addition to introductory statistics classes. Using real and relevant data gives students valuable practice with the data management skills needed to prepare data for statistical analyses (Carver & Stephens 2014). The relevance of the dataset can be particularly meaningful here: if the variables are measures common to the students' discipline, they may have an understanding of normal clinical ranges for lab values or the importance of particular validated scales. With the Licorice Gargle dataset, students may need to create new variables for analysis. For instance, if the smoking variable codes (1=current, 2=past, 3=never) were used directly in analysis, the results would not have a meaningful interpretation. In contrast, dummy variables for "current" and "past" levels can be created with "never" as a reference to arrive at an interpretable model. Working with real data provides students practice with reading a data dictionary, codebook, or equivalent, as well as with data management, e.g., importing data, cleaning data, merging datasets, and creating new variables. More generally, students acquire an appreciation of the challenges, time, effort, and decisions associated with preparing analytic datasets.

Real data, typically, confronts students with the challenges of missing data; the Licorice Gargle dataset is missing outcomes for two patients. Requiring students to grapple with identifying, characterizing, and coding missing data sensitizes them to the importance of preventing missingness and the impact of missing data on subsequent analyses. Finally, exercises in data management confer the additional advantage of introducing students to the importance of transparent and reproducible methods of data collection and management. In our view, the classroom setting is an ideal (and safe!) venue for students to learn the basics of data management and, importantly, to experience the consequences of making mistakes.

Presentation

As, ultimately, the value of statistics lies in its communication to others, the ability to understand statistical results in the literature ("statistical literacy") and to clearly present statistical results to others (presentation skills) are well-recognized essential aims of introductory statistics (Parke 2008). However, in our experience, exercises in the communication of statistics are often overlooked in introductory statistics courses, likely due to time constraints. We have found that the use of real and relevant data provides ready and easy opportunities for communication skills development. For example, as part of the study introduction for a real and relevant dataset, teachers and students can critique the presentation of results in the study publication. Students can then use the data to recreate the tables and figures, thus revealing the connection between the data and the presentation. Additionally, students can be asked to interpret the figures and tables to answer the hypotheses under study. Revealing the story behind data that are presented in tables and figures may not be intuitive for all learners and is generally a skill that must be practiced.

Threading

Using real data with relevant context and background requires more time upfront to explore and understand. "Threading", an often-used strategy in teaching, provides an opportunity to "reuse" a dataset; the time for its introduction is thus freed up for other uses. Threading begins with the teacher communicating a base concept or experience, in order to provide a foundational understanding. On subsequent occasions, the teacher revisits and expands upon this base, which both reinforces the foundational understanding and communicates a new and more advanced understanding. Real and relevant data are ideally suited to this didactic approach. The Licorice Gargle dataset is one such "rich" dataset; it contains multiple variables of various types, thus lending itself easily to being revisited at multiple points in a course. For example, the Licorice Gargle dataset might be used first for descriptive statistics, then for hypothesis testing of categorical data, and still later for logistic regression analysis of binary and potentially ordinal outcomes. Threading may also lessen the fear of handling real data, often termed "analysis paralysis," by desensitizing students through repeated exposure over the course of the semester. Threading also permits a more realistic view of data analyses from start to finish, since the teacher can demonstrate that real projects require descriptive statistics as well as a number of different types of bivariate tests and/or multivariable models conducted on the same dataset. Indeed, it is rare that a single chi-square test, for example, is sufficient for a complete analysis, as our current methods of instruction often imply.

Confidence and Transference

The use of real and relevant data, particularly if more than one dataset is used, gives the student an increased level of confidence that they will be able to transfer their newly learned skills to the data they will encounter in their careers. If the only data they have encountered are "toy" datasets, they may reasonably wonder whether what they are learning will be applicable to "real" datasets. When they have used real and relevant data in class, they know that they have developed the skills needed to manage and analyze real data in the future. Similarly, we have observed that when students have spent time learning to read and understand the statistical results in published study articles, they gain a tremendous level of confidence that they will be able to transfer this skill to reading and evaluating the literature in their own field.

Multiple Uses

The multiple benefits of real and relevant data, in turn, suggest multiple uses in teaching. Real and relevant data can be used for in-class examples, as they are more interesting than "toy" or context-free data. Real and relevant data can also be used for in-class analysis projects; these might include having students work with a partner to carry out the analysis that was just discussed. Two added advantages of in-class projects are that they break up an otherwise long lecture and they facilitate an active learning classroom. Assignments to be done outside of class may involve analyzing real and

relevant data and writing an analysis report and/or drafting portions of a manuscript (e.g., methods and results sections). Lastly, real and relevant data can be used on quizzes and exams. When well-documented datasets are available, the possibilities are limitless and we encourage instructors to get creative.

1.2 OPTIMAL CHARACTERISTICS OF REAL AND RELEVANT DATA FOR TEACHING IN THE HEALTH SCIENCES

Not all real and relevant data are suitable for teaching. Here, we highlight the attributes that enhance the instructional suitability of a dataset.

Raw Data

Raw data are generally preferable to summary data. Raw data allow students to reproduce study results and make connections between data and results. While summary statistics and study results are often made available for teaching purposes, this is not the same as making the actual study data itself available. Occasionally one can use simulations to generate plausible data to match the summary statistics. However, it quickly becomes challenging to reconstruct multivariable correlations or situations of confounding. Unfortunately, published study data remain quite rare, in spite of initiatives and incentives for making study data public. Indeed, notwithstanding new requirements from the National Institutes of Health (NIH) to publish data generated by NIH-funded studies (https://grants.nih.gov/policy/sharing.htm), as of this writing, it continues to be difficult to get permission to use research data for educational purposes.

Data with Context

Data with context are preferable to "naked" datasets. The context of a dataset is an essential part of its appeal to students. Questions such as why the study was done, what the research questions were, how the study was done, and what the measured variables were, are key to piquing the students' interest in this particular set of data and are essential for developing a full understanding of the fundamental process of scientific research. Ideally the background information will include a citation to the published study.

Topical Data

Topical data are generally preferable to general-interest data. Selecting a variety of datasets that are timely and that pertain to the unique topics of interest to the specific group of students can reasonably be expected to pique students' curiosity. While classic datasets are available on Olympic medal counts and survival of Titanic passengers, statistics students who are medical or public health professionals are much more likely to be intrigued by public health inquiries, such as studies of post-operative complication counts or survival following an Ebola outbreak. These students, particularly those who are graduate or professional students, are especially focused on what they need to be successful in their careers; as such they may be bored and possibly annoyed with what they may see as "childish" and irrelevant examples from sports or fashion or other non-medical areas. When a variety of topical datasets are presented, students learn to transfer their knowledge across health contexts. Topical datasets offer the additional benefit to medical and public health professionals of providing opportunities to hone their skills in subject matter analysis and interpretation.

Current Data

Current or controversial data are generally preferable to historic data. Over time and with overuse, even historic, classic datasets can become outdated or boring. An example is the Framingham dataset (see https://biolincc.nhlbi.nih.gov/teaching/). The Framingham study, which began in 1948 and is ongoing with the third generation of participants, has provided important information about the epidemiology of cardiovascular and metabolic diseases (Hajar 2016). The associated teaching dataset, however, is limited to data from the 1950's through the 1970's. Understandably, today's students may not feel it is reflective of modern medical care or lifestyles. Using a current dataset is more likely to facilitate demonstration of current, cutting edge medical or public health research, and the associated development of new statistical methodology. Since many students envision themselves participating in medical and public health research upon graduation, analyses of modern topics provide an intrinsic motivation to learn relevant statistical methodology.

Rich Data

Rich datasets are preferable to minimal datasets. While there is utility in small, simple datasets that can be easily handled by novice students at the beginning of their first course, ultimately, rich, complex datasets that are suitable for a variety of analyses can play a larger role in learning. As discussed in section 1.1, rich datasets support threading, that is, being able to use the same dataset repeatedly through the course to address multiple research questions of various types. Also as noted previously, rich data are more likely to support the full spectrum of statistics from discussions of study design, testable hypotheses, and data management, to analyses, presentation of results, and assessment of conclusions.

We propose a hierarchy of desirable characteristics of real and relevant datasets for teaching in the health sciences, shown in Figure 2. In this hierarchy, datasets that possess more of the desirable characteristics are judged to be of more use to teachers. A real dataset by itself is of limited utility. The addition of basic documentation consisting of a data dictionary and a description of the study makes the dataset much more useful. Adding the published manuscript makes it more useful still, and adding complete documentation including the study protocol and annotated case report forms makes it the most useful (and rarest) of all.

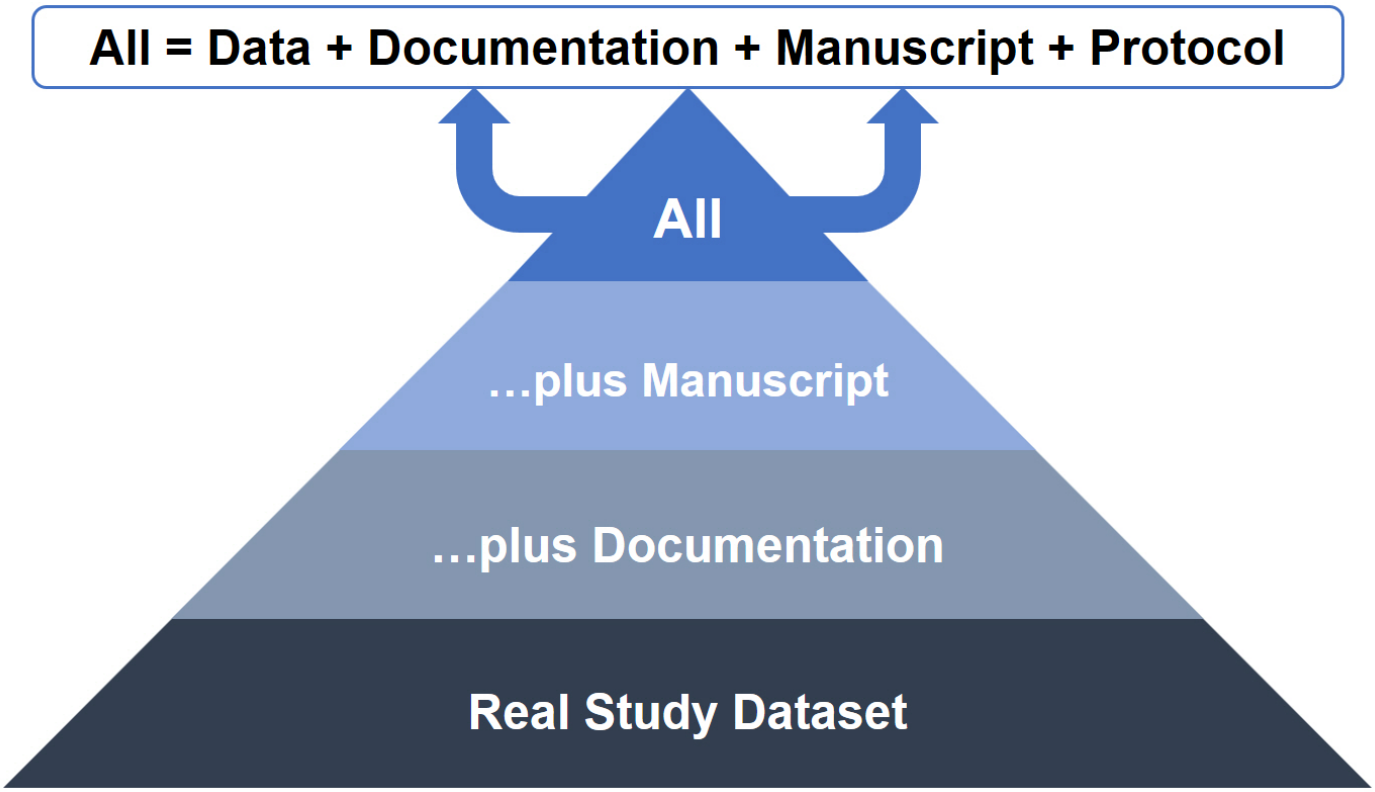

Figure 2: Hierarchy of desirable characteristics for real and relevant health sciences teaching datasets.

2. CHALLENGES IN USING REAL AND RELEVANT DATA

Unfortunately, datasets at the top of the hierarchy of desirable characteristics (Figure 2) are rarely obtainable. In their absence, instructors continue to use datasets that are less than ideal. Here, we consider the barriers to finding and utilizing real and relevant data.

2.1. REAL AND RELEVANT DATA ARE DIFFICULT TO FIND

Instructors often opt not to use real and relevant data for a number of reasons including, especially, that it is difficult and time-consuming to find, particularly in the health sciences. There are many places to find such data, but they are widely scattered, not always well known and of varying degrees of utility.

Textbooks
Some statistics textbooks include data CDs or links to the publisher websites, and some may even print relatively simple datasets in the text. These datasets may have limited context information or may not be "rich", thus placing them in the lowest level of the hierarchy of "good" health sciences datasets (Figure 2).

Journals
A few journals support online supplements or require that authors make their data available through other means, e.g., public repositories, personal research web pages, or by contacting the author (Taichman et al. 2017). In some older pre-HIPAA publications, the entire dataset may be printed out in the article itself. Unfortunately, readers cannot query articles based on the accessibility of data; thus, their use entails spending impracticable amounts of time perusing the scientific literature. An alternative approach would be to "cold call" the authors with requests for the data but, in our experience, this is rarely successful.

Collaborators
Occasionally the instructor's own research collaborations yield intriguing data of value to teaching. However, typically, collaborators are reluctant to allow their data to be used for teaching, especially when the research is still ongoing or is as yet unpublished.

Online
Real data is available in abundance on websites and in research repositories. However, access is not straightforward and many challenges exist. For example, education-focused sites often contain little health-related data. By comparison, research-focused sites rarely contain datasets in forms suitable for teaching and often specifically disallow the use of their datasets for teaching. Some online sites routinely post interesting summary data and graphics, but these sites typically do not permit downloading the associated source datasets (e.g., clinicaltrials.gov). Furthermore, the related primary publications (if any) are not always cited. Finally, online resources are only useful if they can be found. Study-related websites are not often maintained after the associated grant funding ends, leading to frustration due to dead links.

Students
The requirement that students obtain or bring their own data for use in a course is an appealing strategy for piquing student interest but may not be practical in all cases. Graduate students working on their dissertation project may have their own research data, which is naturally highly motivating, but many students of statistics are not at that point in their studies yet. The datasets that students are able to obtain or find may not be appropriate to the course objectives, may be limited in other ways and, possibly, may not meet human subjects protection requirements.

Thus, while there are a multitude of sources for real and relevant data, they are often prohibitively time consuming to find and impractically limited in the resources that are offered.

2.2 REAL AND RELEVANT DATA ARE CHALLENGING TO USE

The lucky teacher who has found "good" real and relevant data is now confronted with the task of preparing it for use. Substantial challenges arise here, too.

Size
Real and relevant datasets may be too small or too large to be useful in class. Very small datasets may be less interesting to students, less "real" seeming, and less useful for revisiting throughout the semester. Here we note that such datasets

do lend themselves to discussions of power and generalizability. Very large datasets (e.g., NHANES survey data) may be less interesting because of their sheer size; everything is significant but nothing is clinically meaningful. While this too provides the opportunity to discuss right-sizing a study, the production of subsets, either population specific or randomly drawn, may help to extract a manageable and analyzable dataset from a massive source. Further, some massive datasets, such as those from continuous monitoring devices or high-throughput assays, may be too large for analysis on a personal computer.

Messiness

Real and relevant datasets may also be too clean or too messy for the instructor's purposes. It is difficult to find datasets with just the right amount of missing data or patterns of missingness for efficient instruction. Ideally, the data should be messy enough to require some thought by the student, e.g., detecting an outlier or creating a new variable, but not so messy so as to be overwhelming, e.g., requiring advanced programming skills.

Limited data

So-called "limited" teaching datasets may be obscured, simplified, or truncated (in observations or in variables) so that it is not possible for students to replicate the published results or answer questions of interest.

Topic suitability

Many datasets suitable for teaching in general are not suited for teaching in the health sciences, because they do not contain medical or public health data. Similarly, it is sometimes difficult to find datasets that are relevant and interesting to students in a particular health-related specialty (e.g., physicians) that are not so discipline-specific that students in other specialties (e.g., dentists) cannot effectively engage with them. This challenge is particularly vexing in introductory classes comprised of students from multiple specialty areas.

Documentation

In our experience, very few datasets are accompanied by a data dictionary. Indeed, typically, a description of the study purpose and design, the observations and variables, and the source of the data is limited or altogether lacking. Also extremely rare is a dataset that is accompanied by the study protocol or case report forms.

Classroom Utility

Since published data are research-derived, they cannot reasonably be expected to be accompanied by teaching materials (e.g., lecture slides, homework or project assignments, learning activities, or quiz questions). Accordingly, with research-based real data, it may not be clear what analysis questions to pose to a class of statistics learners. On this point we note that, as teachers, we often have specific concepts that we want our students to "discover" through analysis, e.g., confounding, mediation, or interaction. Success in this regard thus requires additional effort on the instructor's part, first to explore the dataset's utility in teaching a particular concept and, second, to develop an appropriate experience for the students.

2.3 REAL AND RELEVANT DATA MAY REQUIRE A PARADIGM SHIFT

There are also challenges involved in using real and relevant data beyond finding and preparing the dataset itself. In our view, fully integrating real and relevant data into the learning experience may require a shift in the learning environment.

Limited classroom time

Limited classroom time can make it difficult to cover the statistical concepts and methods, the context of the real data being used, and the software skills needed to carry out the analyses within a single session. Furthermore, analyzing real data will take more time than analyzing simplified or "toy" data due to the potential need for data cleaning and to false starts in the analysis process. One approach to obtaining more time is to fully or partially invert (or "flip") the classroom (Winquist & Carlson 2014, McGraw & Chandler 2015, Schwartz et al. 2016). This involves requiring students to do specified work before coming to class; e.g., reading the textbook, listening to recorded lectures, and/or completing a "pre-class" quiz. Formal class time is then freed up and can be devoted to application and data analysis. Another

approach to obtaining more time is to revisit the same dataset more than once in the same course (see again, Section 1.1, threading).

Varied student backgrounds

Variations in student background make it challenging for the instructor to provide a good learning experience for everyone, particularly in introductory courses. Students who are practicing health care professionals may already have taken some statistics courses or may already have been involved in research. Students who are undergraduates or entry-level graduate students may have no statistics background and no exposure to research. Similarly, some students may have a lot of facility with computer use and even programming, while others, such as older, nontraditional students returning to the classroom, may not. Working with real and relevant data during class time tends to exacerbate this diversity in skill sets. Working in teams, providing additional challenge questions for advanced learners within a rich dataset, or employing additional instructor or teaching assistant time may help to address the diversity while minimizing frustration for the students.

*Faculty incentive*s

An indirect challenge for instructors is that there is very little incentive for faculty or staff instructors to spend time on finding and preparing real and relevant data for teaching. While teaching evaluation scores may play a role in the promotion or tenure application package, it is usually grants, publications, and recognition among peers that drive the evaluation process.

Thus, instructors seeking to use real and relevant data would benefit from a ready collection of datasets, each with documentation and associated teaching materials, and collectively and centrally located in a simple website having a distinct health science focus. If this data resource could additionally support the academic path of instructors who used it, that would be an added benefit.

3. THE TSHS RESOURCES PORTAL

To address the need for a well-designed repository of real and relevant data, the Teaching of Statistics in the Health Sciences (TSHS) section of the American Statistical Association has developed the TSHS Resources Portal. The mission of the TSHS Resources Portal (hereafter referred to as the Portal) is to promote excellence in the teaching of statistics in the health sciences through the dissemination of peer-reviewed datasets and accompanying teaching materials that are centrally archived in an easily-navigated public domain website (https://www.causeweb.org/tshs/category/dataset/).

The Portal is data-centered. Each dataset contribution has its own webpage that houses the data together with all accompanying materials, including documentation and teaching materials. The Portal is also designed to be a dynamic community-oriented endeavor. Contributors submit datasets and associated teaching materials, which undergo a rigorous peer-review process. They are then made available for use in teaching, to share with fellow teachers, and to be further developed by users providing suggestions for additional resources.

The Portal was intentionally designed to provide datasets with the optimal characteristics described earlier (Section 1.2) for teaching in the health sciences. All Portal datasets are health-sciences focused. All datasets provide the raw data. Nearly all datasets are rich, real data from published studies and are accompanied by information on the context. The Portal also includes a few simulated datasets which are valuable for illustrating specific concepts (e.g., methods for correlated data). Nearly all datasets are from recently published studies. Furthermore, the Portal was intentionally designed so that the datasets are near the top of the hierarchy of usefulness described earlier (Figure 2). All Portal datasets are accompanied at minimum by basic documentation: data dictionary, study introduction, and, very often, a reference to the published study.

3.1 WHAT POTENTIAL USERS NEED TO KNOW

As of this writing, there are eleven datasets available on the TSHS Resources Portal. They represent all stages of biomedical research, from discovery, to translation, to clinical application. The Tumor Growth Dataset "contains repeated measurements of tumor size from an animal xenograft experiment designed to compare four treatments for cancers" (Daskalakis 2016). The Season Effect Dataset contains data from an observational study assessing whether season of the year was associated with the likelihood of post-surgical infection (Ngendahimana 2016). The Licorice Gargle Dataset, mentioned earlier, contains data from a randomized controlled trial comparing the effectiveness of a licorice gargle with a sugar solution gargle for the control of post-extubation coughing and sore throat for patients undergoing tracheal tube intubation prior to elective surgery (Nowacki 2017).

Teaching resources are also available for some datasets. These may include lecture slides, tutorials, homework assignments, projects, test questions, clicker questions, applets and more. To facilitate the process of identifying useful teaching resources, each item is accompanied by a Teaching Resource Overview containing a description of the intended audience, the resource type, the learning objectives or goals and a list of all files needed to utilize the resource. All teaching resources are conveniently located on the same webpage with the associated dataset.

Several other features of the Portal are noteworthy. First, each dataset is available for download in multiple formats: R, SAS, STATA, SPSS, Minitab and Excel. Second, each dataset contains supporting documentation in a standardized format. Specifically, there is a "Data Dictionary" which lists the name, label, units and coding for each variable in the dataset, and a two to three page "Dataset Introduction" which describes the study background, objective, design, subjects and variables, and provides a citation for the published study. Both the data dictionary and the dataset introduction are provided in PDF format, suitable for sharing directly with students. Third, each dataset undergoes peer review by the Editorial Board before posting. The review ensures completeness, consistency and high quality. Fourth, for each dataset, any accompanying teaching materials similarly undergo a peer-review process. Finally, the Portal is free and online, easily accessible by all.

The Portal was developed, designed, and intended to be utilized by all levels of statistics educators. This includes K-12 instructors laying the foundations of uncertainty, variability, and data exploration; undergraduate instructors formalizing how uncertainty is quantified; and graduate instructors revealing the power of statistics to address complicated real-world problems. Health sciences students will appreciate the specific biomedical focus aligning with their chosen career paths. Yet all students, regardless of level, have been to a doctor/dentist or visited a hospital; thus, all are able to appreciate the health services sector.

The Portal was designed to serve both instructors and students. However, it is recognized that instructors and students have distinct needs and that the needs of one cannot come at the expense of the other. In particular, exam questions and solutions become unusable to instructors if students can openly access them. Therefore, the Portal has two levels of use: open-access and membership. "Open access" visitors have access to all datasets; no login is required. "Membership" users are teachers; these users have additional access privileges, namely to the posted solutions accompanying teaching materials (note – all teaching materials posted on the Portal include solutions). Membership is offered to all instructors and signing up is both free and simple, requiring only the instructor's name, academic affiliation, and email address. A member of the TSHS Resources Portal Editorial Board verifies instructor status and grants membership (typically within 24 – 48 hours). Upon receiving membership status and an accompanying login code, membership users (teachers) can login anytime to any and all of the resources available: datasets, teaching resources, and solutions to teaching resources.

3.2 WHAT POTENTIAL CONTRIBUTORS NEED TO KNOW

Because the TSHS Resources Portal is data-centric, it accepts three kinds of resources: (1) a new dataset; (2) a new dataset with accompanying teaching resource(s); or (3) a new teaching resource to accompany an existing Portal dataset. Extensive consideration and testing has gone into the submission process so as to render it simple, fast and functional. Detailed instructions are provided on the "For Contributors" tab of the Portal website (https://www.causeweb.org/tshs/contributors/).

The growth and sustainability of the Portal relies on contributions from interested parties such as you. Our hope is that users of the Portal will find the resources extremely useful, saving energy, funds and time. We also hope that regular users will not only reap the benefits of the shared resources but will also be inclined to contribute. So why contribute? One important benefit to contribution is that all posted materials undergo a peer-review process by the TSHS Resources Portal Editorial Board. Therefore, accepted materials may be cited on curriculum vitae. In addition, all resources are properly credited to the contributing author so the Portal serves as a great venue for increasing exposure and reputation. Finally, some grants or funding agencies require developed materials to be disseminated upon completion and the Portal is an excellent option for achieving such a goal.

3.3 CONCLUSION

The TSHS Resources Portal has been designed as an evolving entity, constantly growing and diversifying. With its standardized submission formats and peer-review process, the Portal provides real and relevant data with a level of quality assurance unobtainable elsewhere. It is driven by users and contributors, so be sure to check it out, contribute, and keep coming back to find additional time-saving resources.

4. ACKNOWLEDGEMENTS

This paper grew out of a topic-contributed panel discussion, "Using 'Real Data' for Teaching in the Health Sciences: Benefits, Challenges, and Opportunities" (Session #127), held at the Joint Statistical Meetings (JSM) in Seattle in August 2015.

We would like to acknowledge all who contributed time and energy to developing the TSHS Resources Portal. The organizing committee was led by Carol Bigelow (University of Massachusetts, Amherst), and included Ann Brearley (University of Minnesota), Constantine Daskalakis (Thomas Jefferson University), Deborah Dawson (University of Iowa), Felicity Enders (Mayo Clinic), Edward Gracely (Drexel University), Steven Grambow (Duke University), Jodi Lapidus (Oregon Health Sciences University), Amy Nowacki (Cleveland Clinic), Robert Oster (University of Alabama, Birmingham), Roslyn Stone (University of Pittsburgh), and Susan Telke (University of Minnesota).

The TSHS Resources Portal is hosted by CAUSEweb (www.causeweb.org). We thank Dennis Pearl, director of the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE), as well as Kathy Smith and Bob Casey (all at Pennsylvania State University), and Justin Slauson (Ohio State University), for their invaluable guidance and support in the development and maintenance of the Portal.

We would also like to thank the US Conference on Teaching Statistics (USCOTS) 2013, the Biometrics Section of the American Statistical Association (ASA) and the ASA Section on Teaching of Statistics in the Health Sciences, who provided funding or in-kind resources to support the development of the TSHS Resources Portal.

5. REFERENCES

Briggs, S. (2014), "How to Make Learning Relevant to Your Students (And Why It's Crucial to Their Success)," https://www.opencolleges.edu.au/informed/features/how-to-make-learning-relevant/

Broman, K.W. and Woo, K.H. (2017), "Data organization in spreadsheets," *The American Statistician.* https://doi.org/10.1080/00031305.2017.1375989

Carver, R., and Stephens, M. (2014), "It is Time to Include Data Management in Introductory Statistics," in Makar, K., de Sousa, B., and Gould, R. (Eds.), *Sustainability in Statistics Education: Proceedings of the Ninth International*

Conference on Teaching Statistics (ICOTS9, July 2014), Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute, available at http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_C134_CARVER.pdf

Daskalakis, C. (2016), "Tumor Growth Dataset", *TSHS Resources Portal*. https://www.causeweb.org/tshs/tumor-growth/

Enders, F.T., Lindsell, C.J., Welty, L.J., Benn, E.K.T., Perkins, S.M., Mayo, M.S., and Oster, R. A. (2017), "Statistical competencies for medical research learners: What is fundamental?", *Journal of Clinical and Translational Science*, 1–7. https://doi.org/10.1017/cts.2016.31

Everson, M. (co-chair), Mocko, M. (co-chair), Carver, R., Gabrosek, J., Horton, N., Lock, R., Rossman, A., Rowell, G.H., Velleman, P., Witmer, J., Wood, B. (2016), "Guidelines for Assessment and Instruction in Statistics Education College Report 2016", the American Statistical Association. http://amstat.org/asa/education/Guidelines-for-Assessment-and-Instruction-in-Statistics-Education-Reports.aspx.

Garfield, J. (chair), Aliaga, M., Cobb, G., Cuff, C., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, B., Utts, J., Velleman, P., Witmer, J. (2005), "Guidelines for Assessment and Instruction in Statistics Education College Report", the American Statistical Association. http://amstat.org/asa/education/Guidelines-for-Assessment-and-Instruction-in-Statistics-Education-Reports.aspx.

Garfield J., and Ben-Zvi, D. (2009), "Helping students develop statistical reasoning: Implementing a statistical reasoning learning environment," *Teaching Statistics*, 31(3), 72–77.

Hajar, R. (2016), "Framingham Contribution to Cardiovascular Disease," *Heart Views: The Official Journal of the Gulf Heart Association*, 17(2), 78–81. http://doi.org/10.4103/1995-705X.185130

Libman, Z. (2010), "Integrating Real - Life Data Analysis in Teaching Descriptive Statistics: A Constructivist Approach," *Journal of Statistics Education*, 18(1). http://ww2.amstat.org/publications/jse/v18n1/libman.pdf

McGraw, J.B., and Chandler, J.L. (2015), "Flipping the Biostatistics Classroom, With a Twist," *Bulletin of the Ecological Society of America*, 96(April), 375–384. https://doi.org/10.1890/0012-9623-96.2.375

Neumann, D.L., Hood, M., and Neumann, M.M. (2013), "Using Real-life Data when Teaching Statistics: Student Perceptions of this Strategy in an Introductory Statistics Course," *Statistics Education Research Journal,* 12(2), 59-70. https://iase-web.org/documents/SERJ/SERJ12(2)_Neumann.pdf .

Ngendahimana, D. (2016), "Season Effect Dataset", *TSHS Resources Portal*. https://www.causeweb.org/tshs/season-effect/

Nowacki, A.S. (2017), "Licorice Gargle Dataset", *TSHS Resources Portal*. https://www.causeweb.org/tshs/licorice-gargle/.

Parke, C.S. (2008), "Reasoning and Communicating in the Language of Statistics," *Journal of Statistics Education,* 16(1). http://ww2.amstat.org/publications/jse/v16n1/parke.html

Ruetzler, K., Fleck, M., Nabecker, S., Pinter, K., Landskron, G., Lassnigg, A., You, J., and Sessler, D.I. (2013), "A Randomized, Double-Blind Comparison of Licorice Versus Sugar-Water Gargle for Prevention of Postoperative Sore Throat and Postextubation Coughing," *Anesthesia Analgesia,* 117: 614 – 21.

Samsa, G.P., Lee, L.S., and Neal, E.M. (2012), "An Active Learning Approach to Teach Advanced Multi-predictor Modeling Concepts to Clinicians," *Journal of Statistics Education,* 20(1), 1–34.

Scheaffer, R.L. (2001), "Statistics education: Perusing the past, embracing the present, and charting the future," *Newsletter for the Section on Statistical Education,* 7(1). https://higherlogicdownload.s3.amazonaws.com/AMSTAT/56c109df-f1a1-45e2-bb2b-a6880c16f76a/UploadedImages/v7n1.html#Perusing

Schulz, K.F., Altman, D.G., Moher, D., CONSORT Group (2010), "CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials," *PLoS Medicine,* 7(3): e1000251.

Schwartz, T.A., Andridge, R.R., Sainani, K.L., Stangl, D.K., & Neely, M.L. (2016), "Diverse perspectives on a flipped biostatistics classroom," *Journal of Statistics Education*, 24(2), 74–84. https://doi.org/10.1080/10691898.2016.1192362

Smith, A.E. (2008), "Techniques in Teaching Statistics: Linking Research Production and Research Use," *Journal of Public Affairs Education,* 18(1): 107 – 136. http://www.naspaa.org/JPAEMessenger/Article/VOL18-1/08_smithmartinez-moyano.pdf

Taichman, D.B. et al. (2017), "Data Sharing Statements for Clinical Trials: A Requirement of the International Committee of Medical Journal Editors," *Annals of Internal Medicine,* 167(1): 63-65. http://doi:10.7326/M17-1028.

Willett, J.B., and Singer, J.D. (1992), "Providing a *Statistical* 'Model': Teaching Applied *Statistics* using Real-World Data," in Gordon, F. and Gordon, S. (eds.) *Statistics for the Twenty-first Century*, Washington, DC: Mathematical Association of America, MAA Notes, 26: 83-98.

Winquist, J.R., and Carlson, K.A. (2014), "Flipped Statistics Class Results: Better Performance Than Lecture Over One Year Later," *Journal of Statistics Education*, 22(3), 1–10. https://doi.org/10.1080/10691898.2014.11889717