Data sharing, a four week overview with a critical review of existing practices

### Project Summary/Abstract, 30 lines of text

Not available yet.

### Project Narrative, Three sentences

Not available yet.

### Specific Aims, 1 page

See the separate file, available in [html format](http://www.pmean.com/post/r25-specific-aims.html) or [R Markdown code](https://github.com/pmean/peer-reviewed-data-sets/blob/master/src/r25-specific-aims.Rmd).

### Research Education Program Plan (25 pages)

#### Introduction

The lack of reproducibility in research is a serious problem with serious consequences. Nowhere is this more obvious than in a study by Dr. Anil Potti.

In 2006, Dr. Potti led a team of researchers at Duke University and the University of Sourth Florida to develop and publish information about a gene expression profiles that could predict the sensitivity of cancer patients to a broad range of chemotherapeutic drugs (Potti 2006). They used the term “signature” in the title and repeatedly in the paper itself to emphasize that this profile provided a level of individuality comparable to the name that you use to sign your checks and contracts. Further work expanded this approach to other drugs.

A pair of Statisticians, Keith Baggerly and Kevin Coombes, reviewed the data sets associated with these publication and found numerous errors (Baggerly 2009). The most telling is for one study where “poor documentation led a report on drug A to include a heatmap for drug B and a gene list for drug C.”

This work eventually led to complete invalidation of the model and the withdrawal of the paper. This was more than just an academic exercise. Results from these flawed studies led to investment in a start-up company and the initiation of several clinical trials.

This was not just a “one off” event. A careful review of 18 studies published in Nature Genetics found that that 8 had serious problems with data (Ioannidis 2009). Either the data was not available, it was changed after publication, it was available only in summary form, or it lacked proper documentation.

Reproducibility is “the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator” (Bollen 2015). Reproducibility is a priority area at the US National Institutes of Health (Collins 2014). It is supported by prominent researchers in Clinical Trials research (Vickers 2006), Computational Science (Davison 2019), Ecology (Powers 2019), Epidemiology (Peng 2006), Social, Behavioral, and Economic Science (Bollen 2015) and many other areas.

An external assessment of reproducibility requires that the original researchers share their research data with others. The International Committee of Medical Journal Editors, the Naitonal Institutes of Health, and the National Science Foundation, all have mandates for data sharing. These mandates recognize the importance of patient privacy and the need for the original research team to get their publication out first. Nevertheless, data sharing is expected to be the norm.

Unfortunately, data sharing is hindered by the frequent use of bad storage formats. Researchers often rely on binary formats like Microsoft Word or Adobe PDF that are difficult to import into most statistical software. Another binary format, Microsoft Excel, is easily importable, but suffers from many data quality issues like the conversion of a large number of gene names (Ziemann 2016).

Data documentation can be incomplete or sometimes totally missing. This was an issue in the data shared by Andrew Vickers in his call for greater data sharing (Vickers 2006). His data sets, included with the article, used number codes 0 and 1 for gender without any documentation as to which of these numbers represented females.

#### Proposed Research Education Program

We propose a curriculum on data sharing that covers the mandate for data sharing, the ethical and privacy concerns, and the importance of complete documentation. The material would be accessible to trainees and investigators at any career level. You will be able to follow this material without any advanced of specialized skills in computer science or statistics. Students would only need the ability to produce written documents with any reasonable word processing software and the calculate simple descriptive statistics like means, standard deviations, and percentages using spreadsheet software or something similar.

The material will be broad in scope and applicable to a wide range of scientific and medical disciplines. The only requirement is that the discipline in question regularly produces quantitative research and publishes this research in peer-reviewed journals.

We propose a four week format for this curriculum. This allows time for practice in many of the skills needed for developing a solid data sharing plan. You can use this as a stand alone course for a single hour of academic credit. You can also incorporate this material into an existing course on research methodology or ethics.

Week 1 will cover the mandate for data sharing. This mandate comes from granting agencies (National Institutes of Health 2018, National Science Foundation 2011, Bill and Melinda Gates Foundation, Date Unknown). It also comes from journals (Science 2017, Elsevier 2017). The recent COVID-19 outbreak has led to a pledge to share all relevant research data (Wellcome Trust 2020). Some of these mandates are actually closer to recommendations and all of them allow for important exceptions such as protecting patient privacy.

There is also a strong case for data sharing that relies on basic ethical principles. You do not “own” the research data that you collect. It is owned by the research subjects and is donated to you with the understanding that you will use it to help society. Sharing that data (subject to appropriate pricacy restrictions) is one of the ways that you can insure that this donation has an impact.

One of the exercises in this module will require students to summarize the data sharing mandates for five journals of their own choosing. The students will identify how many of these mandates are actually just recommendations and will note what exceptions each journal offers to the sharing of data.

Week 2 will cover privacy issues. Privacy issues are important both in their own right and in how they complicate the task of data sharing. This starts with a review of the Health Information and Patient Accountability Act (HIPPA). Genetic data requires special review because confidentiality concerns extend beyond the patient to close family members. There are common data security practices that you should always follow with any data that is confidential. You can also de-identify data through deletion of sensitive fields, making random shifts in date fields, adding small amounts of noise to sensitive variables, and ensuring that all crosstabulations have a minimum cell size.

One of the exercises in this module will ask students to identify a published guide on ethical standards in a medical or scientific discipline of their choice. They would then summarize those componenets of that ethical standard that relate to privacy.

Week 3 will cover data documentation. This will start with the FAIR principles that requires data to be findable, accessible, interoperable, and reusable (Wilkinson 2016). This would emphasize why you want your data to be machine-readable and human-readable (Starr 2015). The materials would stress the importance of data standards such as the minimum information about a microarray experiment (Brazma 2001) but also note that many data sets cannot easily follow a specific data standard. This requires the use of general purpose data repositories, such as Dryad (Greenberg 2009). We will define what metadata is and why it is important. We explain the various open source licenses used for many data sets and what would be legally allowed under the fair use provisions of copyright law.

#### Table 1. Topics covered during the first three weeks.

* Week 1. The mandate for data sharing
  + Requirements from granting agencies
  + Requirements from journal editors
  + The ethical case for data sharing
  + Overcoming economic and social barriers to data sharing
* Week 2. Privacy concerns
  + HIPAA
  + Special issues with genetic data
  + data security practices
  + how to deidentify data
* Week 3. Data documentation
  + FAIR (Findable, Accessible, Interoperable, and Reusable) principles
  + data standards
  + general purpose data repositories
  + quality control
  + data stewardship
  + metadata and annotation
  + interoperability

The fourth week would be devoted to a team project that provides a critical appraisal of the quality of documentation of peer-reviewed data sets from a representative sample of research publications in a particular area. Teams of 3-5 students, under the guidance of a professional librarian would develop a well-documented and reproducible search strategy in an area selected by the team. This could be a search of bibliographic resource such as Pubmed that covers a topic of interest to the team. It could be a hand search of articles in a particular journal. It could search an alternate database such as a clinical trials registry.

The search would produce a comprehensive sampling frame from which a representative sample could be selected. The ideal sample size depends on the amount of effort required to review each research publication and the probability that a publication would have a data set available for review. In most settings, the vast majority of publications will fail to include a data set for review, but this could be assessed very quickly.

For those publications with a data set, the teams would document the size of the data set, the format used, and the quality of the data documentation (if any). Each data set would be independently assessed by at least two students with any discrepancies resolved by the entire team.

The teams would then download the data sets to attempt a partial reproduction of statistics that appear in the paper itself. The team would only be asked to reproduce simple descriptive statistics (means, standard deviations, ranges, percentages) and not the more complex results that require the use of advanced statistical models. The team will document whether their attempt at reproducing these descriptive statistics were successful and whether missing information in the data documentation impeded their efforts.

The teams would then produce a written report that documents the methods they used, summarizes the general findings, and discusses the implications of their findings.

The team project produces several benefits. First, the requirement that the students draw a representative sample would reinforce certain skills such as defining a population, creating a sampling frame, pulling a random sample, extracting information from the sampled publications, and summarizing that information in a report. This is great practice in research methods that does not require any IRB review, since it is using publicly available information. Second, the report produced by the team could be presented at a research conference or submitted to a peer-reviewed journal. Third, the students will long remember the inadequate documentation that slows down or prevents reproduction of the descriptive statistics in a particular journal article. They will know what type of documentation is important for their own research studies and will be more likely to provide that documentation.

#### Table 2. The final project

* Teams of 3-5 students
  + Develop a well documented search strategy
  + Produce a comprehensive sampling frame
  + Select a representative sample of publications/data sets
  + Document documentation quality
  + Attempt to reproduce table of descriptive statistics
  + Summarize findings in a research paper

This curriculum will include closed captioned videos of didactic lectures on data sharing mandates, privacy and ethical concerns, and good data documentation practices. Powerpoint slides with speaker notes will be provided for those who wish to incorporate this material into their own lectures. Discussion questions and practical exercises will emphasize finding, accessing, interpreting, and re-using open data sets.

All materials will be developed using R Markdown, a markup language that uses simple text files, and stored on a github repository. The use of R Markdown and github will allow others to easily adapt and improve the materials. R Markdown has the capability of producing files in a wide range of formats (docx, html, pdf, pptx) which further increases the flexibility of these materials.

Each didactic lecture will be recorded and placed on a YouTube channel for general viewing. The Powerpoint slides along with speaker notes will be available if you wish to prepare your own lectures.

All the materials (the videos, the Powerpoint slides with speaker notes, the homework exercises with grading rubrics, and the details on the final team project) will be arranged on Canvas, a popular learning management system. The entire Canvas site can be archived and made available to other users.

All teaching materials will comply with [section 508 with the Rehabilitation Act](https://www.section508.gov/manage/laws-and-policies) and [Web Content Accessibility Guidelines](https://www.w3.org/WAI/standards-guidelines/wcag/). All non-text content that is presented to the user has a text alternative that serves the equivalent purpose. Videos will include closed-captioning for the hearing impaired as well as the speaker notes in a Powerpoint file. All written materials will be designed to simplify the use of screen-reading sotware. Where color is used, it will not be the sole way to convery information. All materials will use high contrast and allow easy resizing of the content to help those with vision impairment.

These materials will be licensed under the Creative Commons attribution license (CC-BY 4.0). This is the most liberal open source license available and will allow users to create derivative works.

#### Timeline for module development.

To be added.

#### Program Director/Principal Investigator

The administration of this program is truly a team effort. The team has a balanced set of skills in curriculum development, evaluation, library science, and statistics.

The principal investigator, Steve Simon, is a passionate advocate of the use of real world data sets. Every class that he teaches is littered with data sets that come from peer-reviewed publications. The search for these data sets, however, has been filled with difficulties due to limited search tools and poor documentation. He has a vested interest in documenting these difficulties and advocating for improvements.

Dr. Simon is a popular and highly sought after teacher. He has provided short courses at research conferences for Alternative Medicine, Andrology, Medical Librarians, Pediatricians, and Statisticians. The greatest testament to his work is the repeat invitations that he gets from these conferences.

Dr. Simon has extensive experience with remote teaching. He has provided numerous seminars for The Analysis Factor, all conducted via GoToMeeting. He developed a popular multi-week training course on survival analysis using the same platform. All of the classes he teaches at the University of Missouri-Kansas City are run as asynchronous online courses. Three of the online classes were brand new, developed or co-developed from scratch.

Finally, Dr. Simon has already developed and tested much of the material to be developed for this training module on sharing data. He includes videos on data dictionaries and program documentation in his courses on R, SAS, and SQL. He discusses data management at great length in his Clinical Research Methodology class. He has provided guest lectures on data sharing to another faculty member’s course, Responsible Conduct of Research. While the material does need to be consolidated, all the pieces are already sitting out there waiting to be assembled.

List a paragraph with the background and strengths of each of the other members of the research team.

The research team will meet monthly to discuss progress. All materials will be shared using version control software (git), allowing a coordinated approach to the management and update of individual files.

#### Institutional Environment and Commitment

(From the RFA) “Describe the institutional environment, reiterating the availability of facilities and educational resources (described separately under “Facilities & Other Resources”), that can contribute to the planned Research Education Program. Evidence of institutional commitment to the research educational program is required. A letter of institutional commitment must be attached as part of Letters of Support (see below). Appropriate institutional commitment should include the provision of adequate staff, facilities, and educational resources that can contribute to the planned research education program."

#### Evaluation Plan

(From the RFA) “Applications must include a plan for evaluating the activities supported by the award in terms of their frequency of use and their usefulness. The use of multiple evaluation approaches is highly encouraged, as is testing several groups with different characteristics. The application must specify baseline metrics (e.g., numbers, educational levels, and demographic characteristics of test groups) in a structured format, as well as quantifiable measures to gauge the short and long-term success of the research education award in achieving its objectives, including testing for content comprehension and retention. Applicants are expected to obtain feedback from test groups to help identify weaknesses and to provide suggestions for improvements, and make the evaluation and feedback data available to the NIH staff.”

#### Dissemination Plan

(From the RFA) “Applications must include a specific plan for disseminating the finished training modules online and their digital material nationally, and for making them accessible to everyone at no cost. In addition, links to these modules and their digital material will be posted and maintained on the NIGMS clearinghouse web site.”

### Biographical sketch (5 pages)

### Letters of support

[List here the people who I have approached and the ones that have agreed to provide letters of support.]

\*IMPORTANT" (From the RFA) “A letter of institutional commitment must be attached as part of Letters of Support (see section above:”Institutional Environment and Commitment“). Applications that do not include this Letter of Support will be withdrawn prior to review.”

# Everything from this line forward is NOT part of the grant submission

### Link to RFA and other resources.

The grant I am applying for is

National Institutes of Health. RFA-GM-20-001: Training Modules to Enhance the Rigor, Reproducibility and Responsible Conduct of Biomedical Data Science Research. Available in [html format](https://grants.nih.gov/grants/guide/rfa-files/RFA-GM-20-001.html).

This is an R25 grant. The R25 grant is intended for the development of research education activities, as described [here](https://researchtraining.nih.gov/programs/research-education/r25). The page limits for an R25 grant are available [here](https://grants.nih.gov/grants/how-to-apply-application-guide/format-and-write/page-limits.htm).

You can get a great idea of the range of materials already funded by reviewing educational resources posted on the [NIGMS clearinghouse web site](https://www.nigms.nih.gov/training/pages/clearinghouse-for-training-modules-to-enhance-data-reproducibility.aspx).

### Miscellaneous resources

Request for Information (RFI) on Developing an Online Educational Resource for Training in the Principles of Rigorous Research. Notice Number: NOT-NS-20-062. Available in [html format](https://grants.nih.gov/grants/guide/notice-files/NOT-NS-20-062.html).

National Institutes of Health. Training Instructions for NIH and Other PHS Agencies - Forms Version F Series. Available in [pdf format](https://grants.nih.gov/grants/how-to-apply-application-guide/forms-f/training-forms-f.pdf).

General application guide for NIH and other PHS Agencies. SF424 (R&R) - Forms Version F. Available in [html format](https://grants.nih.gov/grants/how-to-apply-application-guide/forms-f/general/g.100-how-to-use-the-application-instructions.htm) or [pdf format](https://grants.nih.gov/grants/how-to-apply-application-guide/forms-f/general-forms-f.pdf).

#### Important quotes from the RFA

(From the RFA) “State the goals for education and justify the area of training selected for module development in terms of its relevance and potential impact on improving the development of skills and knowledge important for rigorous, reproducible, and responsible conduct of biomedical data science research. Describe the data science subject material to be covered, and provide details of the evidence-based methods that will be employed to achieve the training goals. Describe the format for the proposed training module and justify it in terms of the education goals. Chosen subject material may address a particular level of technical complexity/expertise, but the proposed research education must target a broad audience, i.e., trainees and investigators at any career level. The length of the proposed training module should be explained in terms of scope and depth of coverage of the subject matter. In addition, how the research education will be utilized by trainees or investigators should be described - for example, a module on how to properly handle biomedical data to be taken by all beginning laboratory workers, or a module on appropriate data provenance to be taken immediately prior to beginning such work. Describe the plans for piloting and evaluating the effectiveness of the training module. Describe plans for making the proposed training module section 508 compliant with the Rehabilitation Act (29 U.S.C. ’794 d), as amended by the Workforce Investment Act of 1998 (P.L. 105 – 220; see <http://www.section508.gov/> for additional information). Describe the electronic format of the proposed modules and discuss how the used formats are standard, making themodules widely shareable. Provide a timeline for module development, piloting and refinement, dissemination, evaluation, and maintenance. This timeline must propose making the training publicly available within two years of the award date.”

(From the RFA) “Describe arrangements for administration of the program. Provide evidence that the Program Director/Principal Investigator is actively engaged in research and/or teaching in an area related to the mission of NIH, and can organize, administer, monitor, and evaluate the research education program. For programs proposing multiple PDs/PIs, describe the complementary and integrated expertise of the PDs/PIs,their leadership approach, and governance appropriate for the planned project.”

exportable training modules designed to enhance the rigor, reproducibility, and responsible conduct of biomedical and behavioral data science research, targeted to trainees and researchers at any career level.

rigor and reproducibility issues can also stem from how biomedical data is handled and processed, such as poor curation or missing provenance information or lack of appropriate quality controls during data processing. Rigor and reproducibility problems may also arise from unintentional mishandling of the data sets due to lack of standards, open file formats, or interoperability of data sets.

data privacy and security, appropriate use of informed consent, ethics of handling personal data, potential for bias, and data sharing on a need-to-know basis.

This new FOA proposes to expand such efforts to address specifically the training needs in areas relevant to the rigor, reproducibility, and responsible conduct of biomedical data science research.

enhance the scientific rigor, reproducibility, and responsible conduct of biomedical data science research

relatively short units of training of sufficient depth and coverage to empower the trainee with the knowledge and skills to improve the rigor, reproducibility, and responsible conduct of their data science research

shareable and accessible, open educational resources, which can be used as standalone modules or combined to form courses

any stage of the data lifecycle (from data acquisition, to organizing, filtering, annotating, mining, visualizing, publishing, and/or preserving data) and data issues related to data stewardship, quality control, ethics, and regulatory policies

the proposed training modules will identify deficiencies and teach best practices in one or more data science areas

Scientific principles of rigorous and reproducible data science Good practices in scientific rigor and reproducible data science Good practices in the responsible conduct of data science research

The training modules are expected to cover material not typically taught as part of current institutional coursework

It is expected that all modules will be developed, piloted, and disseminated within the first two years of the award; budgets may be awarded for up to three years to allow module evaluation and module modification/maintenance in year three

### Evaluation criteria from the RFA

#### Overall Impact

Reviewers will provide an overall impact score to reflect their assessment of the likelihood for the project to strongly advance research education by fulfilling the goal of this R25 Education Program, in consideration of the following review criteria and additional review criteria, as applicable for the project proposed.

#### Scored Review Criteria

Reviewers will consider each of the review criteria below in the determination of scientific merit, and give a separate score for each. An application does not need to be strong in all categories to be judged likely to have major scientific impact.

#### Significance

Does the proposed program address a key audience and an important aspect or important training need in rigor, reproducibility, and responsible conduct of data science research Is there convincing evidence in the application that the proposed program will significantly advance the stated goal of the program?

#### Investigator(s)

Is the PD/PI capable of providing both administrative and scientific leadership to the development and implementation of the proposed program? Is there evidence that an appropriate level of effort will be devoted by the program leadership to ensure the program’s intended goal is accomplished? If applicable, is there evidence that the participating faculty have experience in mentoring students and teaching science? If applicable, are the faculty good role models for the participants by nature of their scientific accomplishments? If the project is collaborative or multi-PD/PI, do the investigators have complementary and integrated expertise; are their leadership approach, governance and organizational structure appropriate for the project?

#### Innovation

Taking into consideration the nature of the proposed research education program, does the applicant make a strong case for this program effectively reaching an audience in need of the program’s offerings? Where appropriate, is the proposed program developing or utilizing innovative approaches and latest best practices to improve the knowledge and/or skills of the intended audience?

#### Approach

Does the proposed program clearly state its goals and objectives, including the educational level of the audience to be reached, the content to be conveyed, and the intended outcome? Is there evidence that the program is based on a sound rationale, as well as sound educational concepts and principles? Is the plan for evaluation sound and likely to provide information on the effectiveness of the program?

Are the proposed modules and their digital material in standard and shareable format, and easily available to the public at no cost?

#### Environment

Will the scientific and educational environment of the proposed program contribute to its intended goals? Is there a plan to take advantage of this environment to enhance the educational value of the program? Is there tangible evidence of institutional commitment? Where appropriate, is there evidence of collaboration and buy-in among participating programs, departments, and institutions?

### Additional thoughts (things that I’ve written but I’m not ready to place them in a particular spot in the grant)

One attempt to reproduce 18 published studies on gene expression failed completely for 10 of those studies and could only produce a complete reproduction for 2 of the remaining 8 (Ioannidis 2009).

Outline of materials coverd

Great quote:

“Metadata is a love note to the future” Jason Scott, on a [tweet dated 2011-09-29](https://twitter.com/textfiles/status/119403173436850176?lang=en).

In 2006, Anil Potti led a team of researchers at Duke University and the University of Sourth Florida to develop and publish information about a gene expression profiles that could predict the sensitivity of cancer patients to a broad range of chemotherapeutic drugs. They used the term “signature” in the title and repeatedly in the paper itself to emphasize that this profile provided a level of individuality comparable to the name that you use to sign your checks and contracts. Further work expanded this approach to other drugs.

A pair of Statisticians, Keith Baggerly and Kevin Coombes, reviewed the data sets associated with these publication and found numerous errors. The most telling is for one study where “poor documentation led a report on drug A to include a heatmap for drug B and a gene list for drug C.”

This work eventually led to complete invalidation of the model and the withdrawal of the paper. This was more than just an academic exercise. Results from these flawed studies led to investment in a start-up company and the initiation of several clinical trials.

This was not just a “one off” event. A careful review of 18 studies published in a journal that required sharing of the raw data sets found that that 8 had serious problems with data. Either the data was not available, it was changed after publication, it was available only in summary form, or it lacked proper documentation.

Reproducibility is “the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator” (Bollen 2015). Data sharing is a critical component to insuring that research is reproducible.

### Early versions of specific aims page

Note: a more recent version of the specific aims page is found in [html format](http://www.pmean.com/post/r25-specific-aims.html) or [R Markdown code](https://github.com/pmean/peer-reviewed-data-sets/blob/master/src/r25-specific-aims.Rmd). I am includeing versions 1 and 2 of the specific aims page because they were more detailed (and unfortunately well beyond the limit of a single page). I may borrow some material from these longer aims to insert into the grant itself.

### Version 1 of specific aims page

If you think that data sharing provides an independent avenue for assessing research reproducibility, think again. While there are many mandates for data sharing, the mandates remain largely unfulfilled (Vickers 2016). If you use open data in the wrong way, you can be characterized as a “research parasite” (Longo 2016). You will find the reproduction of others work is tedious and labor intensive, due to poor documentation (Baggerly 2009). When you find problems with reproducibility, journal editors make it difficult for you to alert readers to these problems (Allison 2016). If you want any hope for reproducibility of research results by independent parties, then data sharing must become a research norm and the data that is shared must be done with appropriate documentation.

#### Long term goal

Our long term goal is to enhance the scientific rigor, reproducibility, and responsible conduct of biomedical data science research through a four week curriculum on data sharing that you can use in a stand-alone one credit hour class or which you can incorporate into an existing course on research ethics or methodology. This curriculum will include closed captioned videos of didactic lectures on data sharing mandates, privacy and ethical concerns, and good data documentation practices. Powerpoint slides with speaker notes will be provided for those who wish to incorporate this material into their own lectures. Discussion questions and practical exercises will emphasize finding, accessing, interpreting, and re-using open data sets. The work will culminate in a team project where students conduct a critical review of representative sample of open data sets in a particular scientific or medical specialty. This review will include an attempt to replicate the descriptive statistics typically found in the first table of most research publications (e.g., age range, percent female).

#### Aim 1. Pilot testing of team project.

The team project is the heart and soul of the training module. Reproducing the complex statistical models in most research papers would require too much time as well as specialized expertise beyond the skill set of most students. But a reproduction of simple descriptive statistics, such as means, standard deviations, and percentages should certainly be possible. We plan a series of pilot studies with small teams of students in four disciplines: Biology, Economics, Medicine, and Psychology. These teams will be asked to identify twelve data sets associated with peer-reviewed publications in a particular area of their choosing. They will then attempt to reproduce at least one descriptive statistic from the associated paper. They will be provided with written resources on data repositories, and search strategies, and will track the amount of time needed to find the data sets as well as the amount of time to produce and check at least one descriptive statistic. This could be anything, but we want to encourage the use of something simple like an age range or percentage female that typically would appear in the first table of most students involving humans. The students will then provide suggestions on the written resources that we provided and review the features of the data sets that made it easy or difficult to reproduce some simple descriptive statistic.

The pilot study will help estimate the difficulty of the critical review and help scale the problem up or down so that it is challenging but not excessively burdensome. It will also help us prepare a more refined set of instructions and informational materials.

#### Aim 2. Development and review of training materials.

The research team will develop Powerpoint slides and a script included in the speaker notes covering three major areas: the mandate for data sharing, privacy and ethical concerns, and data documentation standards. We will also develop exercises for each area along with grading rubrics. We will seek external reviewers for this material and revise the materials as needed.

The materials will form the foundation for an existing class on Clinical Research Methodology, but also be available in raw form for anyone who wants modify the materials to meet the particular needs of their teaching environment.

#### Aim 3. Deploy materials in Clinical Research Methodology class and conduct evaluation.

Videos with closed captioning will be prepared for use in the Clinical Research Methodology class taught by the principal investigator. The Canvas learning management system will provide access to the videos and exercises. Students will provide informal evaluations of the materials and be tested on their retention of key concepts immediately afte the material and again at the end of the class. Students who provide consent will be contacted six months later for an additional assessment.

These assessments will be made public (after appropriate anonymization of indivdual responses) to allow us and other parties to revise and improve the training materials.

#### Aim 4. Make all resources will be made available publicly.

All the Powerpoint slides, speaker notes, handouts, exercises, and grading rubrics will be developed using RMarkdown and stored on a github site. The videos produced will be posted on a Youtube channel.

All products produced by this grant will be licensed under the [Creative Commons Attribution license](https://creativecommons.org/licenses/by/4.0/) (CC-BY 4.0). This is the most liberal license available and will allow users to create derivative works.

#### Summary

Data sharing is a vital step to insure reproducibility of research findings by external parties. We propose a four week overview of data sharing that culminates in a critical review of a representative sample of open data sets. This material is not typically taught as part of current coursework on research ethics or research methdology. These materials will help raise the level of knowledge about the practices of data sharing and help turn an unfulfilled mandate into a research norm.

### Version 2 of specific aims page

#### Aim 1. Pilot testing of team project.

The team project is the heart and soul of the training module. Reproducing the complex statistical models in most research papers would require too much time as well as specialized expertise beyond the skill set of most students. But a reproduction of simple descriptive statistics, such as means, standard deviations, and percentages should certainly be possible. We plan a series of pilot studies with small teams of students in four disciplines: Biology, Economics, Medicine, and Psychology. These teams will be asked to identify twelve data sets associated with peer-reviewed publications in a particular area of their choosing. They will then attempt to reproduce at least one descriptive statistic from the associated paper. They will be provided with written resources on data repositories, and search strategies, and will track the amount of time needed to find the data sets as well as the amount of time to produce and check at least one descriptive statistic. This could be anything, but we want to encourage the use of something simple like an age range or percentage female that typically would appear in the first table of most students involving humans. The students will then provide suggestions on the written resources that we provided and review the features of the data sets that made it easy or difficult to reproduce some simple descriptive statistic.

The pilot study will help estimate the difficulty of the critical review and help scale the problem up or down so that it is challenging but not excessively burdensome. It will also help us prepare a more refined set of instructions and informational materials.

#### Aim 2. Development and review of training materials.

The research team will develop Powerpoint slides and a script included in the speaker notes covering three major areas: the mandate for data sharing, privacy and ethical concerns, and data documentation standards. We will also develop exercises for each area along with grading rubrics. We will seek external reviewers for this material and revise the materials as needed.

The materials will form the foundation for an existing class on Clinical Research Methodology, but also be available in raw form for anyone who wants modify the materials to meet the particular needs of their teaching environment.

#### Aim 3. Deploy materials in Clinical Research Methodology class and conduct evaluation.

Videos with closed captioning will be prepared for use in the Clinical Research Methodology class taught by the principal investigator. The Canvas learning management system will provide access to the videos and exercises. Students will provide informal evaluations of the materials and be tested on their retention of key concepts immediately afte the material and again at the end of the class. Students who provide consent will be contacted six months later for an additional assessment.

These assessments will be made public (after appropriate anonymization of indivdual responses) to allow us and other parties to revise and improve the training materials.

#### Aim 4. Make all resources will be made available publicly.

All the Powerpoint slides, speaker notes, handouts, exercises, and grading rubrics will be developed using RMarkdown and stored on a github site. The videos produced will be posted on a Youtube channel.

All products produced by this grant will be licensed under the [Creative Commons Attribution license](https://creativecommons.org/licenses/by/4.0/) (CC-BY 4.0). This is the most liberal license available and will allow users to create derivative works.

### Where to find this document and related documents.

* Bibliography in [html format](http://www.pmean.com/post/r25-bibliography.html) or in [R Markdown code](https://github.com/pmean/peer-reviewed-data-sets/blob/master/src/r25-bibliography.Rmd).
* R25 grant draft in [html format](http://www.pmean.com/post/r25-grant.html) or in [R Markdown code](https://github.com/pmean/peer-reviewed-data-sets/blob/master/src/r25-grant.Rmd).
* An earlier brief overview of the R25 grant in [html format](http://www.pmean.com/post/training-module-summary.html) or in [R Markdown code](https://github.com/pmean/peer-reviewed-data-sets/blob/master/src/training-module-summary.Rmd).
* Draft publication on figshare in [html format](http://www.pmean.com/post/data-sharing-publication.html) or in [R Markdown code](https://github.com/pmean/peer-reviewed-data-sets/blob/master/src/data-sharing-publication.Rmd).
* R25 specific aims in [html format](http://www.pmean.com/post/r25-specific-aims.html) or [R Markdown code](https://github.com/pmean/peer-reviewed-data-sets/blob/master/src/r25-specific-aims.Rmd).