ISDS 2019 Conference Abstracts

Examining and improving reproducible research practices in public health

Kimberly J. Johnson, Bobbi J. Carothers, Xiaoyan Wang, Todd Combs, Douglas A. Luke, Jenine K.

Harris Public Health, Washington University in St. Louis, St. Louis, Missouri, United States

Objective

Our presentation will explain current use, and barriers to use, of reproducible research practices in public health. We will also introduce a set of modules for researchers wishing to increase their use of reproducible research practices.

Introduction

An important goal of surveillance is to inform public health interventions that aim to reduce the burden of disease in the population. Ensuring accuracy of results is paramount to achieving this goal. However, science is currently facing a "reproducibility crisis" where researchers have found it difficult or impossible to reproduce study results. Organized and well-documented statistical source code that is publicly available could increase research reproducibility, especially for research relying on publicly available surveillance data like the BRFSS, NHANES, GSS, SEER, and others. As part of our overall goal to improve training around reproducible research practices, we surveyed public health data analysts to determine current practices and barriers to code sharing.

Methods

We conducted a cross-sectional web-based survey about code organization, documenting, storage, and sharing. We surveyed public health scientists who reported recently conducting statistical analyses for a report or manuscript. A total of 247 of 278 screened eligible to filled out the survey, and 209 answered every applicable question. We used traditional descriptive statistics and graphs to examine the survey data.

Results

Most participants reported using some promising coding practices, with 67% including a prolog to introduce the code and 85% including comments in statistical code to explain operations and analyses. Of 10 common code organization strategies (e.g., naming variables logically, using white space), most (82%) respondents reported employing at least three of the strategies and just under half (47%) reported using five or more. Over half of participants (59%) reported code was developed or checked by two or more people. Many participants also reported promising file management habits for data and code used in publications. Three- quarters (75%) had a variable dictionary to accompany the dataset used, 48% created clean versions of code files, and 64% created clean versions of data files at the time of publication. Forty three percent of participants reported that if they suddenly left their current position, it would not be easy for others to find their statistical code files. Public code sharing was much less common among participants with just 9% reporting sharing code publicly from a recent publication and 20% of those surveyed reported ever having shared code publicly.

The top two barriers to using reproducible research practices were lack of training in reproducible research (n=108) and data privacy issues (n=105). Journals and funders not requiring reproducible practices were barriers selected by 94 and 84 participants, respectively. Few participants identified fear of errors being discovered (n=26) or a lack of workplace incentives (n=32) as barriers.

Conclusions

Most participants were using some promising practices for organizing and formatting statistical code but few were sharing statistical code publicly. The second most frequently identified barrier to using reproduciible practices was data privacy, which could prohibit easily sharing a data source. With surveillance data often being publicly available, researchers working with surveillance data have overcome this top barrier without any change to current research practices.

Researchers using surveillance data could greatly increase research reproduciblity by adopting promising practices for code formatting, like using logical variable names and limiting line length, and posting code in a public repository like GitHub.

To overcome the top barrier to use of reproducible research practices, lack of training, we developed brief training modules on formatting, documenting, and sharing statistical code and data. As part of our presentation we will introduce and provide access to these online modules. The introduction will focus on the relevant modules for surveillance data users, which include statistical code formatting and statistical code sharing via GitHub.

With fewer barriers to practicing reproducible research, public health researchers using surveillance data have the opportunity to be leaders in improving the adoption of reproducible research practices and subsequently improving the quality of research we rely on to improve public health.

Acknowledgement

This project was supported by the Robert Wood Johnson Foundation (RWJF) Increasing Openness and Transparency in Research program.