Conference Materials 2017

Cynthia M. Kroeger, Andrew W. Brown, David B. Allison
5/9/2018

Project Description

This project was implemented to further develop methods that were originated previously *(insert DOI for 2016 abstract)* as part of a larger project (F32DK107157) to study common biases and statistical mistakes made in nutrition and obesity research. One such bias is p-hacking. As the poster describes in more detail, this project replicated a previously implemented rapid and high-volume method for detecting p-hacking in scientific literature. This work has since been further developed and presented on, and a manuscript with fully-developed methods is now being prepared for publication.

Conference Description

This work was presented at the 5th Workshop on Biostatistics and Bioinformatics in Atlanta, Georgia, USA, in 2017.

Contact Information for Investigators

Principal Investigator

- Cynthia M. Kroeger, PhD, Postdoctoral Fellow, Indiana University School of Public Health-Bloomington, Department of Epidemiology and Biostatistics, kroeger@iu.edu

Co-investigators

- Andrew W. Brown, PhD, Assistant Professor, Indiana University School of Public Health-Bloomington, Department of Applied Health Sciences, awb1@iu.edu
- David B. Allison, PhD, Dean, Distinguished Professor, Provost Professor, Indiana University School of Public Health-Bloomington, Department of Epidemiology and Biostatistics, allison@iu.edu

At the time of presentation, Dr. Kroeger was a Postdoctoral Fellow at the University of Alabama at Birmingham, and Drs. Brown and Allison were affiliated with the University of Alabama at Birmingham.

Discription of Supporting Materials

Presentation: Poster

This work was presented as a poster.

File name for poster: conference_materials_2017_poster.pdf

Data

File name for dataset: conference_materials_2017_data.csv

Full names and definitions of column headings

Column Heading	Definition	Possible Values
trial	study trial	org = original trial; rep = replication study
method	search string method	a = method a; b = method b (methods are described in section "Methodological details" below)
database	database data were collected from	pm = PubMed; dss = PubMed Dietary Supplement Subset
date	date data were collected	
mode	mode of data extraction	man = manual extraction
three	number of abstracts containing a 0.03* p-value	

four	number of abstracts containing a 0.04* p-value	
five	number of abstracts containing a 0.05* p-value	
six	number of abstracts containing a 0.06* p-value	
common	number of abstracts containing a common statistical analysis	
atypical	number of abstracts containing an atypical analysis term	
three_exclusive	number of abstracts containing a 0.03* p-value and not a 0.04*, 0.05* or 0.06* p-value	
four_exclusive	number of abstracts containing a 0.04* p-value and not a 0.03*, 0.05* or 0.06* p-value	
	number of abstracts	

five_exclusive	containing a 0.05* p-value and not a 0.03*, 0.04* or 0.06* p-value	
six_exclusive	number of abstracts containing a 0.06* p-value and not a 0.03*, 0.04* or 0.05* p-value	
common_exclusive	number of abstracts containing a common statistical analysis term and not an atypical statistical analysis term	
atypical_exclusive	number of abstracts containing an atypical statistical analysis term and not a common statistical analysis term	
three_common	number of abstracts containing a three_exclusive and common_exclusive term	
	number of abstracts containing a	

four_common	four_exclusive and common_exclusive term	
five_common	number of abstracts containing a five_exclusive and common_exclusive term	
six_common	number of abstracts containing a six_exclusive and common_exclusive term	
three_atypical	number of abstracts containing a three_exclusive and atypical_exclusive term	
four_atypical	number of abstracts containing a four_exclusive and atypical_exclusive term	
five_atypical	number of abstracts containing a five_exclusive and atypical_exclusive	

	term	
six_atypical	number of abstracts containing a six_exclusive and atypical_exclusive term	
n_common	sum of rows columns 6:9	
n_atypical	sum of rows columns 10:13	

These data were collected and formatted by Dr. Kroeger.

Methodological details

Search string details for method a

```
# p-value bin 0.03: (p=.03*[tiab] OR p=0.03*[tiab])
# p-value bin 0.04: (p=.04*[tiab] OR p=0.04*[tiab])
# p-value bin 0.05: (p 05,05[tiab] OR p 050[tiab] OR p 05
# p-value bin 0.06: (p=.06*[tiab] OR p=0.06*[tiab])
# common tests: (t-test[tiab] OR anova[tiab] OR ancova[ti
# atypical tests: (nonparametric[tiab] OR non-parametric[
```

Asterisks in PubMed are used to indicate wildcards. PubMed returns wildcards and shows the terms returned for manual inspection. Expansion of the 0.05* term above shows these wildcards expanded. Search 0.05* is expanded to remove inclusion of the commonly found mention of $p < 0.05$ in abstracts. PubMed interprets any

equality characters (<, =, >) as blank spaces. An asterisk alone could not be used here, as we needed to manually remove references to 0.05. These are not likely to represent values in a bin, because they could represent the entire inequality of $p < 0.05$ or reference to a significance level threshold.

Search string details for method b

```
# p-value bin 0.03: same as method a
# p-value bin 0.04: same as method a
# p-value bin 0.05: same as method a
# p-value bin 0.06: same as method a
# common tests: (t-test[tiab] OR "t test" OR t-student[ti
# atypical tests: (nonparametric[tiab] OR non-parametric[
```

Reasons for changes:

- Common tests: "t test", t-student, ANOVA, and "parametric tests" were added as they are revelant common tests, and different versions (i.e. addition or removal of a "-") yielded additional results.
- Atypical tests: "transformation" was removed, as it was found to yield numerous irrelevant results. "non parametric", wilcoxon-mann-whitney, mann-whitney, u-test, wilcoxon, log-transformed, and "log transformed" were added, as these terms were found to yield additional relevant results.

PubMed query data file

The .csv file with the search query conducted in and exported by PubMed could not be located for this study. Instead, notes were

located on the the search query obtained for the 2016 abstract. Details on this search query can be found here: *(insert 2016 DOI)*. Screenshots of chi-squared tests were located to obtain the data used for this poster and replicate analyses in R for open sharing. To better document the research process and enhance replicability, R and rEntrez were employed to conduct future studies on this topic. Files for this work will be available in the study repository upon completion.

Analysis code

File name for analysis code: conference_materials_2017_code.R

Study level naming convention: database_method_trial_mode

- database: pm = PubMed; dss = PubMed Dietary Supplement Subset
- a = method a; b = method b
- trial: org = original; rep = replication
- mode: man = manual search

Analysis level naming convention: study_function

- study: pm_a_rep_man, dss_a_rep_man, pm_b_rep_man, dss_b_rep_man
- function: xsq = chisq.test; pd = percentage deviation

Code was written by Dr. Kroeger, with assistance from Dr. Brown.

A Note on How to Replicate this Study

Because data were extracted from PubMed manually on a specific date, one would need to download PubMed history and assess the literature from the point in time noted in the data file, using the methods described in this document. Another option might be to run the search described herein and subtract dates from the date reported in our data file. We have not done this ourselves. However, this work does replicate previous findings in the same literature set at a later date.

Author Contribution

Dr. Kroeger conducted the study, refined the methods, analyzed data, and made the poster. Dr. Brown provided editorial assistance with methods and poster preparation. Dr. Allison provided editorial assistance and is the senior author.

License Information

File name for analysis code: LICENSE.txt

Citation Information

Recommended citation for conference_materials_2017

Please use the following to cite any of the supporting materials

herein:

Kroeger CM, Brown AB, Allison DB. Use of text-mining and comparative analysis to detect p-hacking in dietary supplement scientific literature. Zenodo. 2018. DOI: *(insert DOI)*.

Recommended citation for published abstract

Please use the following to cite the published abstract:

Kroeger CM, Brown AW, Allison DB. Use of text-mining and comparative analysis to detect p-hacking in dietary supplement scientific literature. 5th Workshop on Biostatistics and Bioinformatics. Atlanta, Georgia, USA. 2017.

Funding Information

The presentation and development of these materials was supported by the National Institutes of Health (F32DK107157). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or any other organization.