Research Ideas and Outcomes 1: e7462
doi: 10.3897/rio.1.e7462

OPEN ACCESS

Small Grant Proposal

Problematic trial detection in ClinicalTrials.gov

Chris HJ Hartgerink‡, Stephen L George§

‡ Tilburg University, Tilburg, Netherlands
§ Duke University, Durham, NC, United States of America

Corresponding author: Chris HJ Hartgerink (c.hartgerink@gmail.com)

Reviewable v1

Received: 08 Dec 2015 | Published: 17 Dec 2015

Citation: Hartgerink C, George S (2015) Problematic trial detection in ClinicalTrials.gov. Research Ideas and Outcomes 1: e7462. doi: 10.3897/rio.1.e7462

Executive summary

Clinical trials are crucial in determining the effectiveness of treatments and directly affect clinical and policy decisions. These decisions are undermined if the data are problematic due to data fabrication or other errors. Researchers have worked on developing statistical methods to detect problematic data. This project aims to develop new methods and apply them to results reported in the ClinicalTrials.gov database. Using both established and and the newly developed statistical methods we will investigate the prevalence of problematic data, trends of problematic data over time, and whether the prevalence of problematic data is predicted by trial characteristics such as funding type.

Keywords

clinicaltrials.gov, problematic data, error, data fabrication

Background

Clinical trials are crucial in determining the effectiveness of medical interventions and directly affect the clinical and policy decisions made by medical doctors and governmental institutions, respectively. If the clinical trial data are problematic, these decisions are undermined. A salient example is the case where beta-blockers were prescribed to decrease perioperative mortality in cardiac patients, based on what later appeared to be

problematic data (Commissie Vervolgonderzoek 2012). A subsequent meta-analysis indicated beta-blockers actually increased perioperative mortality if the results of the problematic studies were excluded (Bouri et al. 2014). As a result, patients were exposed to increased risk instead of decreased risk; had the problematic data not been detected, they might still have been exposed to increased risk today.

As a result, some researchers have worked on developing statistical methods to detect problematic data, but this remains a niche field despite the large scientific and societal benefits of improving these methods. For clinical trials conducted across different locations, central statistical monitoring can be applied to detect aberrant data by comparing the summary results from different collection sites (Timmermans et al. 2015). Such a monitoring tool is useful because fully verifying the raw data for all locations is practically infeasible (George and Buyse 2015). Others have developed a posteriori methods, which apply summary results or raw data to detect problematic data (reviewed in Buyse et al. 1999). For example, 168 trials by Fuji were analyzed based on published summary results, which indicated that these trials were problematic (Carlisle 2012, Carlisle et al. 2015) and subsequently resulted in more than 100 retractions.

Problematic data can arise from data fabrication, but can also be the result of erroneous procedures or data handling. For instance, imagine a randomized clinical trial that shows severely different baseline measures for two randomized groups. Since these groups are expected to be statistically similar, if severe differences do occur, this can indicate something went wrong in the randomization procedure, measurements were improper for one condition, or data were fabricated (Al-Marzouki et al. 2005).

It remains unclear how prevalent problematic data is throughout the different sciences and how the prevalence has developed throughout the years. Approximately 2% of all researchers admit to having fabricated data, the most egregious form of problematic data, at least once (Fanelli 2009). Since the 2% estimate is based on self-reports, it is plausible that it underestimates the true extent of the problem due to response bias. Anecdotally, the Journal of Cell Biology found that image manipulation, another form of problematic data, was present in 1% of their accepted manuscripts (Journal of Cell Biology 2015).

With many high-exposure cases of problematic data throughout recent years and in different scientific fields (e.g., the Stapel case in psychology, the Fuji case in anesthesiology, the Poldermans case in medicine), it might appear that more problematic data are published now than a decade ago. However, systematic data on how prevalent problematic data are and how the prevalence has developed over time are largely absent. Extending the set of statistical methods available to investigate problematic data helps increase its detection; systematic application of these methods would provide a prevalence estimate of problematic data.

Systematic application of statistical methods to detect problematic data and testing hypotheses of its prevalence could be done within the ClinicalTrials.gov (CT) database, which includes over 100,000 completed trials with summary results for more than 15,000 clinical trials. The CT database, where information on clinical trials are available from

beginning to completion, allows for a substantial amount of meta-research and includes data that are reported in a standardized format. Such meta-research can range from meta-analyses to the adherence of preregistration protocols in publications (i.e., outcome reporting bias). Since, as noted above, an estimated 2% of researchers have admitted to fabricating data at least once, some of these results can be expected (in part) to be based on problematic data. Research related to the ClinicalTrials.gov database has amassed 138 research publications in Web of Science since its launch in 2008 (as of October 5, 2015), but none of these publications explicitly inspected potentially problematic data.

One category of studies available within the CT database are large, multi-location trials that contain aggregated results based on data collected at multiple locations. As such, if there are many research locations, problematic data by one location might be masked by aggregating and consequently remain undetected. Nonetheless, if there are more research locations, the aggregate data at baseline across randomized groups contains less standard error. Thus, problematic data with influential (i.e., extreme) outlying values, which severely distort the mean would also be more readily detected. Whether problematic data is detected is therefore highly dependent on how the data is problematic (e.g., missing values coded as 999 are included in the analysis for a question ranging from 1-10), a question the CT database cannot answer and this project therefore does not aim to answer.

The CT database also contains other categories and this project investigates the prevalence of problematic data for all categories at the study level. These categories include interventional studies and observational studies, including multi-location and single location studies. The proposed project investigates all categories available and will provide a prevalence estimate for all categories together, and per category separately. The prevalence within different categories of clinical trials can be directly compared, which can serve as an indicator for further research.

The prevalence estimates will also be regressed onto trial characteristics, in order to see whether problematic data is more prevalent for studies that contain certain characteristics. For example, the hypothesis that industry-funded research is more frequently problematic than publicly funded research can be directly tested by estimating whether there is a difference between the prevalence estimates for industry- and publicly funded research. Other trial characteristics will be mapped during the data collection and flagged if of interest.

Objectives

1. What new statistical methods can be developed to detect problematic data based on summary results alone (e.g., means, standard deviations, demographics, test results)?
2. Based on the developed methods, what is the estimated prevalence of potentially problematic data in the ClinicalTrials.gov database?
3. Based on the developed methods, do the prevalence estimates of potentially problematic data in the ClinicalTrials.gov database show a change over time?

4. Based on the developed methods, are prevalence estimates of potentially problematic data in the ClinicalTrials.gov database predicted by study characteristics (e.g., funding type)?

Methodology

Clinical trials aim to make inferences about the effectiveness of a treatment in the population by taking a sample of that population and clinical trials that do not adhere to the principles of sampling theory can be said to be problematic. For example, earlier research highlighted that researchers are often quite bad in fabricating data that look genuinely stochastic (Mosimann et al. 2002, Mosimann et al. 1995). Moreover, if a randomized experimental design includes an honest mistake, for instance assigning all females to the control group, stochasticity does not fully apply anymore.

As a consequence, the principle of stochasticity can be applied to help detect potentially problematic data. Some previous methods have been developed that apply this tenet to detect problematic data. For example, the Simonsohn method (Simonsohn 2013) examines the “variance of variances.” That is, how much the observed variances differ across conditions and whether this is a reasonable amount given natural sampling fluctuations. If the results are highly unlikely under sampling theory (e.g., 10^{-5}), this indicates that something might have gone wrong in the data collection, for whatever reason.

A limited set of methods are currently available to investigate the stochasticity of summary results—methods this project aims to extend. With an extension of these methods, the possibility to detect problematic data based on only summary results could increase. These methods would be tested in simulation studies for their validity before applying them to the CT data themselves. For example, the uniformity of nonsignificant results can be inspected, or the size of effects found in the clinical trials. Such methods can be based on previous misconduct investigations and based on theoretical expectations of the data under sampling theory. For instance, the Fuji case in anesthesiology (Carlisle 2012) used statistical tools to indicate that results were too good to be true. Other cases in different fields also applied statistical tools to detect problematic data, such as the Baltimore case in biology (Research Integrity Adjudications Panel 1996), the Stapel case (Levelt Committee et al. 2012), the Smeesters case, and the Sanna case in psychology (Simonsohn 2013). Considering that methods are developed within the project, the exact methodology is hard to outline a priori. The stochastic tenet will serve as the main driver to develop the methods. The data from the CT database will be extracted and structured prior to the start of the project.

An example of such a method is a χ^2-test for randomized trials, which tests whether a set of baseline demographics data are statistically equal across groups after taking into account that different measures have different expected values. The CT database includes baseline demographics for almost all studies. For each baseline demographic an expected value (i.e., the mean of the randomized groups) and a χ^2-value can be computed.

Repeating this for all reported baseline measures, an overall χ^2-test for potentially problematic data in the baseline is computed as

$$\chi^2_{(M-1)(K-1)} = \sum_{i=1}^{M}\sum_{j=1}^{K}\frac{(O_{ij}-E_i)^2}{E_i}$$

where M is the number of measures and K is the number of groups available. The expected- and observed values should not deviate substantially from each other in randomized trials at the baseline measurement, but could deviate from each other if something problematic occurred. Because this measure looks at all baseline demographics simultaneously, the probability of a positive result due to sample fluctuations (i.e., false positive) decreases.

Significance

Improving the detection of potentially problematic data has two important outcomes: (i) better methods to detect problems before- and after publication of the results and (ii) a new way of estimating the prevalence of problematic data, moving away from biased self-reports. Additionally, the statistical detection tools would be of interest to the Office of Research Integrity in the U.S., the Dutch National Board for Research Integrity (LOWI), academic editors, peer-reviewers, or (potential) whistleblowers, as part of investigating suspected papers. These detection tools might also serve as a deterrent for data fabrication by improving its detection. Moreover, these statistical methods would not be limited to clinical trials, but could also be applied in other fields (e.g., psychology).

Professionally, the project would allow me to learn from one of the few experts in this niche field and discuss the intricacies of researching problematic data. Besides advancing the methods themselves, the limitations of these methods and their ethical implications will provide sufficient discussion during the visit, promoting my academic development with respect to the implications of the methods and their findings. As such, this visit would provide a bridge between my local supervisors (Jelte M. Wicherts and Marcel A.L.M. van Assen), experts located in the Netherlands, and researchers in the U.S., potentially providing further international collaborations in the future.

Evaluation and dissemination

Upon submission, this proposal will be published in the the Research Ideas and Outcomes (RIO) journal for public evaluation. The research project will be publicly documented with the Open Science Framework and GitHub, allowing for direct reproduction of all analyses and results. This allows for rapid dissemination and participation of fellow researchers during the project. The Open Science Framework and GitHub apply version control, which is a track changes for files (Ram 2013), providing a revision history of all changes applied and improves (chronological) documenation of the research process. The statistical methods used during the project will be implemented into the `ddfab` package in R, which

is under development by the applicant and will be made freely available. The results of the project itself will be checked by another researcher (i.e., co-piloted) and written up in a manuscript that will be submitted to a peer-reviewed Open Access journal. This manuscript will be shared as a preprint for public feedback. All research output will be made available with nonrestrictive Creative Commons licenses, with an explicit preference for the public domain (i.e., CC-0) and otherwise only requiring attribution (i.e., CC-BY). Such nonrestrictive licenses allow for maximum re-use and impact of the research.

Justification for residence in the United States for the proposed project

Research on detecting potentially problematic data is a niche field; the number of researchers that also have experience with these methods within the medical sciences is even more limited. Stephen L. George (SLG) combines knowledge of statistics in general, clinical trials, and the application of statistics to detect potentially problematic data due to for example research misconduct. He has extensive knowledge of these methods (Buyse et al. 1999) and how these can be used to provide estimates of incidence or prevalence of research misconduct (George 2015, George and Buyse 2015). Moreover, his thorough case knowledge and mathematical background will help in developing new statistical methods to detect problematic data. Hence, the current project would thoroughly benefit from discussion and supervision by SLG.

Duration

The proposed project lasts six months; Table 1 includes a preliminary specification of activities prior to- and during the visit. The proposed project entails analyzing data retrieved from the ClinicalTrials.gov database; the data will be extracted in the two months prior to visiting the U.S. as preparation. Upon arrival, the applicant outlines the types of data available in the database (including recalculated results) and, together with the onsite supervisor (SLG), inspects which available methods might be useful to apply to this database to detect problematic data. Moreover, upon outlining the types of data available, the applicant and SLG also discuss potential avenues for new methods. The application of these methods to the ClinicalTrials.gov data is postponed until the statistical properties of these methods are further investigated. During this process, the methods are also incorporated as functions in the R statistical environment (R Core Team 2015), if they appear valuable. Subsequently, the remaining methods are applied to the ClinicalTrials.gov database to estimate the prevalence of potentially problematic data.

Table 1.

Preliminary planning of the proposed project, specified per month, including preparatory work prior to visiting.

	2017 (month)							
	Preparation		Visit U.S.					
What	1	2	3	4	5	6	7	8
Data collection from ClinicalTrials.gov	✓	✓						
Recalculate test statistic information		✓	✓					
Review applicability available methods			✓					
Develop new method(s)			✓	✓	✓	✓		
Investigate statistical properties of new method(s)			✓	✓	✓	✓		
Implement methods in R			✓	✓	✓	✓		
Apply methods to data collected from ClinicalTrials.gov						✓	✓	
Write paper on project	✓	✓	✓	✓	✓	✓	✓	✓

English proficiency

The applicant is a proficient English speaker, reader, and writer as a result of his extended exposure to the English language from an early age. His secondary education was in a bilingual fashion (English/Dutch) and he followed an English master programme. These educational aspects helped increase the technical aspects of his English, whereas his vocabulary was built by reading books written or translated by native English speakers.

Funding program

This grant proposal was written for the Dutch Fulbright application (academic year 2016-2017) due on December 1, 2015. This is a direct copy of the submitted research proposal.

Author contributions

CHJH conceptualized the proposal, CHJH drafted the proposal, CHJH and SLG revised the proposal.

References

- Al-Marzouki S, Evans S, Marshall T, Roberts I (2005) Are these data real? Statistical methods for the detection of data fabrication in clinical trials. BMJ 331 (7511): 267-270. DOI: 10.1136/bmj.331.7511.267
- Bouri S, Shun-Shin MJ, Cole GD, Mayet J, Francis DP (2014) Meta-analysis of secure randomised controlled trials of β-blockade to prevent perioperative death in non-cardiac surgery. Heart 100 (6): 456-464. DOI: 10.1136/heartjnl-2013-304262
- Buyse M, George SL, Evans S, Geller NL, Ranstam J, Scherrer B, Lesaffre E, Murray G, Edler L, Hutton J, Colton T, Lachenbruch P, Verma BL (1999) The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. Statistics in medicine 18 (24): 3435-3451. DOI: 10.1002/(SICI)1097-0258(19991230)18:243.0.CO;2-O
- Carlisle JB (2012) The analysis of 168 randomised controlled trials to test data integrity. Anaesthesia 67 (5): 521-537. DOI: 10.1111/j.1365-2044.2012.07128.x
- Carlisle JB, Dexter F, Pandit JJ, Shafer SL, Yentis SM (2015) Calculating the probability of random sampling for continuous variables in submitted or published randomised controlled trials. Anaesthesia 7: 848-858. DOI: 10.1111/anae.13126
- Commissie Vervolgonderzoek (2012) Rapport vervolgonderzoek naar mogelijke schending van de wetenschappelijke integriteit. https://web.archive.org/web/20151027084205/http://www.erasmusmc.nl/5663/135857/3675250/3706798/erasmusmc.commissie.verv.onderzoek.2012
- Fanelli D (2009) How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PloS one 4 (5): e5738. DOI: 10.1371/journal.pone.0005738
- George SL (2015) Research misconduct and data fraud in clinical trials: prevalence and causal factors. International journal of clinical oncology: 1-7. DOI: 10.1007/s10147-015-0887-3
- George SL, Buyse M (2015) Data fraud in clinical trials. Clinical investigation 5 (2): 161-173. DOI: 10.4155/cli.14.116
- Journal of Cell Biology (2015) About the Journal. URL: https://web.archive.org/web/20150911132421/http://jcb.rupress.org/site/misc/about.xhtml
- Levelt Committee, Drenth Committee, Noort Committee (2012) Flawed science: The fraudulent research practices of social psychologist Diederik Stapel. https://www.commissielevelt.nl/
- Mosimann J, Dahlberg J, Davidian N, Krueger J (2002) Terminal digits and the examination of questioned data. Accountability in research 9 (2): 75-92. DOI: 10.1080/08989620212969
- Mosimann JE, Wiseman CV, Edelman RE (1995) Data fabrication: Can people generate random digits? Accountability in research 4 (1): 75-92. DOI: 10.1080/08989629508573866
- Ram K (2013) Git can facilitate greater reproducibility and increased transparency in science. Source code for biology and medicine 8 (1): 7. DOI: 10.1186/1751-0473-8-7
- R Core Team (2015) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing URL: http://www.R-project.org/

- Research Integrity Adjudications Panel (1996) Thereza Imanishi-Kari, Ph.D., DAB No. 1582 (1996). URL: http://web.archive.org/web/20150810164314/http://www.hhs.gov/dab/decisions/dab1582.html
- Simonsohn U (2013) Just post it: The lesson from two cases of fabricated data detected by statistics alone. Psychological science 24 (10): 1875-1888. DOI: 10.1177/0956797613480366
- Timmermans C, Doffagne E, Venet D, Desmet L, Legrand C, Burzykowski T, Buyse M (2015) Statistical monitoring of data quality and consistency in the Stomach Cancer Adjuvant Multi-institutional Trial Group Trial. Gastric cancer: official journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association: 1-7. DOI: 10.1007/s10120-015-0533-9