Ecological performance of detecting data fabrication with summary statistics

Chris HJ Hartgerink, Jelte M Wicherts, Marcel ALM van Assen

February 6, 2017

Any field of empirical inquiry is faced with cases of scientific misconduct at some point, either in the form of fabrication, falsification or plagiarism (FFP). Psychology was faced with Stapel; medical sciences were faced with Poldermans and Macchiarini; life sciences were faced with Voignet. These are just a few examples of cases in the last decade. Overall, an estimated 2% of all scholars have admitted to falsifying or fabricating research results at least once (Fanelli, 2009) and this is likely to be an underestimate due to socially desirable responses. The detection rate is likely to be even lower; for example, only around a dozen cases are discovered in the United States and the Netherlands, despite covering several hundreds of thousands of researchers. At best, this amounts to a detection rate far below 1% of those 2% who admit to fabricating data — the tip of a seemingly much larger iceberg.

In order to stifle attempts at data fabrication, improved detection of fabricated data is considered to deter such harmful attempts. Although deterrence theory dates back to the middle of the 17th century (Hobbes, 1651), its implementation has not occurred across the different forms of scientific misconduct equilaterally. Basically, deterrence theory stipulates that with increased risk of detection, the utility of scientific misconduct (for this context) will decrease and therefore fewer people will engage in such behaviors. This principle of deterrence has been implemented with plagiarism scanners, a development that already started a long time ago (e.g., Parker and et al., 1989). However, increased deterrence of fabrication and falsification by improved detection mechanisms has not been as widely implemented.

In the last decade, detecting image manipulation has become one of the few forms of detecting scientific misconduct other than plagiarism. The Journal of Cell Biology scans each submitted image for potential manipulation (The Journal of Cell Biology, 2015), which greatly increases the risk of detecting (blatant) image manipulation. More recently, algorithms have been developed to automate the scanning of images for (subtle) manipulations (Koppers et al., 2016). These developments in detecting image manipulation have increased detection risk during the pre-publication and post-publication phase by improving detection mechanisms and increasing the understanding of how images might be manipulated. Moreover, their application also helps researchers systematically

evaluate research articles to estimate the extent of the problem of image manipulation (4% of all papers are estimated to contain manipulated images; Bik et al., 2016).

Statistical methods can provide one way to improve detection of data fabrication in empirical research. Humans are notoriously bad at understanding and estimating probabilities (e.g., Tversky and Kahneman, 1974, 1971), which could manifest itself in the fundamentally probabilistic data they try to fabricate. That researchers do not understand probabilistic processes also presents itself in the interpretation of genuine research data (Hoekstra et al., 2006; Sijtsma, 2015; Goodman, 1999). When data are fabricated, probabilistic principles are easily violated if these principles are forgotten at the univariate level, bivariate level, trivatiate level, or beyond (Haldane, 1948). Based on such a theoretical framework, statistical methods that investigate whether the reported data are actually plausible under theoretically probabilistic processes can be used to detect potential data fabrication.

The application of such statistical methods to detect data fabrication has occurred in several cases in recent years and has potential for future application beyond a case-basis. For example, problems with papers by Fuji were highlighted with statistical methods (Carlisle, 2012; Carlisle et al., 2015), resulting in 183 retractions (Oransky, 2015). In this case, baseline measures across randomized groups were examined for too little variation. Random assignment should introduce a certain degree of random error that is might be missed by a human fabricator, misestimating the probabilistic process that generates such error. Another two cases are those of Sanna and Smeesters, where fabricated data were also detected with statistics (Simonsohn, 2013). These cases inspected the variance of variances (i.e., the second level, or meta, variance). Once again, too little variation was what revealed problems in these data. These methods, although developed in a case-setting, need not be limited to cases. The application of such methods can be (semi-)automated if data are available in a machine-readable format that one of the statistical methods can be applied to. An example of such a potential case for mass application of using statistics to detect (potential) data fabrication is in the ClinicalTrials.gov database, where baseline measures across randomized groups are readily available for download and subsequent analysis (Hartgerink and George, 2015).

Nonetheless, considering the potential harm of applying statistical methods to flag potentially problematic results, it needs to be sorted out whether such methods have diagnosticity that actually makes it responsible to apply them. We hardly know how researchers might go about fabricating data. Cases such as Fuji, Smeesters, and Sanna provide some insights, but are highly pre-selected (i.e., those who got caught/confessed) and as such, systematically biased. Relatively extensive descriptions in rare and partial autobiogrophical accounts provide little insight into the actual data fabrication process, except for the setting where it might take place (e.g., late at night when no one is around; Stapel, 2014). Additionally, the performance of methods to detect data fabrication is highly dependent on the unknown prevalence of data fabrication and the power to actually to detect data fabrication. Given that we do not know how

researchers might fabricate data, the diagnosticity of these methods cannot realistically be simulated.

The effectiveness of detection mechanisms and their consequences, hence their expected deterrence, is exacerbated by the increased usage of public online discussion platforms such as PubPeer (`https://pubpeer.com`). PubPeer serves as an "online journal club" where anyone can discuss articles. Authors are notified when someone leaves a response, providing them with the possibility to respond. Such a platform allows for public discussion of the paper, including discussion of reanalyses, methodology, availability of materials, etc. When detection mechanisms are freely available to use, they can lead to (a surge of) comments when applied by users. Recently, in 2016, one user (the main author of this paper) used 'statcheck' software to report potential statistical reporting inconsistencies for 50,000 psychology articles, which led to a large discussion about (automated) online comments (Baker, 2016). The impact of such new possibilities is not to be underestimated, although its potential to contribute to the scientific discussion should also not be. Nonetheless, the 'statcheck' software was well validated prior to this application (Nuijten et al., 2015) and the same principle applies to the application of statistical methods to detect (potential) data fabrication.

Throughout this paper, we inspect statistical methods to detect data fabrication that can be applied to (1) summary results or (2) raw data. Even though the data available look different depending on the structure of the study, there are certain common characteristics of results and the underlying raw data that can be inspected. For example, summary results frequently include means, standard deviations, test-statistics, and p-values. Raw data frequently contain at least some variables measured at a interval- or ratio scale (Stevens, 1946).

We present a set of studies that directly test the validity of statistical methods to detect data fabrication. To this end, Study 1 inspected the performance of statistical methods to detect data fabrication when using only summary results (e.g., means and standard deviations) as typically reported in empirical research articles. Study 2 inspected the performance of statistical methods aimed at detecting data fabrication in raw data (i.e., the data underlying summary results). These two studies provide a first indication of how applicable and effective statistical methods are to detect data fabrication in practice, with actual researchers fabricating actual data.

References

Baker, M. (2016). Stat-checking software stirs up psychology. *Nature*, 540(7631):151–152.

Bik, E. M., Casadevall, A., and Fang, F. C. (2016). The prevalence of inappropriate image duplication in biomedical research publications. *mBio*, 7(3).

Carlisle, J. B. (2012). The analysis of 168 randomised controlled trials to test data integrity. *Anaesthesia*, 67(5):521–537.

Carlisle, J. B., Dexter, F., Pandit, J. J., Shafer, S. L., and Yentis, S. M. (2015). Calculating the probability of random sampling for continuous variables in submitted or published randomised controlled trials. *Anaesthesia*, 70(7):848–858.

Fanelli, D. (2009). How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PloS one*, 4(5):e5738.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine*, 130(12):995.

Haldane, J. B. S. (1948). The faking of genetical results. *Eureka*, 6:21–28.

Hartgerink, C. and George, S. (2015). Problematic trial detection in ClinicalTrials.gov. *Research Ideas and Outcomes*, 1:e7462.

Hobbes, T. (1909/1651). *Leviathan*. Oxford University Presss. http://www.gutenberg.org/ebooks/3207.

Hoekstra, R., Finch, S., Kiers, H. A. L., and Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review*, 13(6):1033–1037.

Koppers, L., Wormer, H., and Ickstadt, K. (2016). Towards a systematic screening tool for quality assurance and semiautomatic fraud detection for images in the life sciences. *Science and Engineering Ethics*, pages 1–16.

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., and Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48(4):1205–1226.

Oransky, I. (2015). The Retraction Watch Leaderboard.

Parker, A. and et al. (1989). Computer algorithms for plagiarism detection.

Sijtsma, K. (2015). Playing with Data-Or How to Discourage Questionable Research Practices and Stimulate Researchers to Do Things Right. *Psychometrika*.

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological science*, 24(10):1875–1888.

Stapel, D. (2014). *Ontsporing [Derailment]*.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684):677–680.

The Journal of Cell Biology (2015). About the Journal. `https://web.archive.org/web/20150911132421/http://jcb.rupress.org/site/misc/about.xhtml`. Accessed: 2015-9-11.

Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76:105–110.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

SessionInfo

```
> sessionInfo()

R version 3.3.2 (2016-10-31)
Platform: x86_64-redhat-linux-gnu (64-bit)
Running under: Fedora 25 (Workstation Edition)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
[1] tools_3.3.2

>
```