Automated detection of data fabrication using statistical tools

Chris HJ Hartgerink, Jan G Voelkel, Jelte M Wicherts, Marcel ALM van Assen

16 April, 2018

Introduction

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 faced Stapel; medical sciences faced Poldermans and Macchiarini; life sciences faced Voignet; physical sciences faced Schön — these are just a few examples of misconduct cases in the last decade. Overall, an estimated 2% of all scholars admit to falsifying or fabricating research results at least once (Fanelli, 2009), which due to its self-report nature is likely to be an underestimate. The detection rate of data fabrication is likely to be even lower; for example, only around a dozen cases become public in the United States and the Netherlands, despite that there are several hundreds of thousands of researchers in these countries. At best, this suggests a detection rate 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 behavior. This idea is based on deterrence theory (Hobbes, 1651), which stipulates that increased risk of detection decreases the utility of scientific misconduct, hence, fewer people will engage in it. Detection techniques have developed to varying degrees for fabrication, falsification, and plagiarism. Plagiarism scanners have been around the longest (e.g., A. Parker & Hamblen, 1989) and are widely implemented not only at journals but also in student evaluation. For data fabrication, developments around detecting image manipulation are more recent and these methods are being implemented at journals. For example, the Journal of Cell Biology and the EMBO journal scan each submitted image for potential manipulation (**???**; The Journal of Cell Biology, 2015), which supposedly increases the risk of (blatant) image manipulation. More recently, algorithms are being developed to automate the scanning of images for such manipulations (Koppers, Wormer, & Ickstadt, 2016). Moreover, the application of such tools can also help researchers systematically evaluate research articles in order to estimate the extent to which image manipulation occurs (4% of all papers are estimated to contain manipulated images, Bik, Casadevall, & Fang, 2016) or what factors are predictive of image manipulation (**???**).

Detection methods for data fabrication in empirical research are often based on a mix of psychology theory and statistics theory. Because humans are notoriously bad at understanding and estimating randomness (**???**; Amos Tversky & Kahneman, 1971; A. Tversky & Kahneman, 1974), this could manifest itself in the fundamentally probabilistic data they try to fabricate. As such, when data are fabricated, principles of statistics and randomness could easily be violated at various dimension of the data (Haldane, 1948). Based on this idea, statistical methods can be used to detect anomalies in investigations. Whether data are in line with the reported probabilistic processes and their theoretically expected outcomes (e.g., random assignment) can indicate deviations from the reported protocol, potentially even data fabrication.

Statistical methods have proven to be of importance in initiating data fabrication investigations or in assessing scope of potential data fabrication. For example, Kranke, Apfel, and Roewer skeptically perceived Fuji's "incredibly nice" data (**???**) and used statistical methods to contextualize their skepticism. At the time, a reviewer perceived them to be on a "crusade against Fujii and his colleagues" (@ **???**) and further investigation was absent. Only when Carlisle extended the systematic investigation to 168 of Fuji's papers (**???**; Carlisle, 2012; Carlisle, Dexter, Pandit, Shafer, & Yentis, 2015) did events cumulate into an investigation- and ultimately retraction of 183 peer-reviewed papers (**???**; Oransky, 2015). In another example, the Stapel case, statistical evaluation of his oeuvre occurred after he had already confessed to fabricating data (Levelt, 2012). This resulted in 58 retractions (Oransky, 2015) and cleared all PhD students of wrongdoing (Levelt, 2012).

In order to determine whether generic application of statistical methods to detect data anomalies is responsible, their diagnostic value requires further investigation. After all, accusations of data fabrication have grave consequences for the people involved (regretfully, the STAP case brings this to the fore very clearly; **???**). Many of these statistical methods to detect data anomalies are quantifications of initial suspicions by researchers. Hence, until validated, these should be considered proposed methods. In some cases, convergent evidence is provided by testing the underlying psychological premises in the general population (**???**, **???**) but the question remains whether this premise also applies for data fabricators. Considering we hardly know how researchers might go about fabricating data, this seems especially problematic. Known cases provide relatively few and biased insights into the mind of the data fabricator. 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).

We present two studies investigating the diagnostic performance of statistical methods to detect data fabrication. These studies investigate methods to detect data fabrication in summary statistics (Study 1) or in raw data (Study 2). In Study 1, we invited researchers to fabricate summary statistics for a set of four anchoring studies, for which we also had genuine data from the Many Labs 1 initiative (https://osf.io/pqf9r; Klein et al., 2014). In Study 2, we invited researchers to fabricate raw data for a Stroop experiment, for which we also had genuine data from the Many Labs 3 initiative (https://osf.io/n8xa7/; **???**). Before presenting these studies, we expand on the theoretical framework of the investigated statistical methods to detect data fabrication.

Theoretical framework

In the current paper, we differentiate between statistical methods to detect potential data fabrication based on reported summary statistics or raw data. Below, we expand on the theoretical underpininings of these methods and provide sample code to run these, where appropriate (implemented in the `ddfab` package for R). For summary statistics, we review *p*-value analysis, variance analysis, and effect size analysis as potential ways to detect data fabrication. *P*-value analyses can be applied whenever a set of nonsignificant *p*-values are reported; variance analysis can be applied whenever a set of variances and accompanying sample sizes are reported for independent, randomly assigned groups; effect size analysis can be used whenever the effect size is reported or can be computed (e.g., an APA reported t- or F-statistic; **???**). For raw data, we review digit analyses (i.e., the Newcomb-Benford law and terminal digit analysis) and multivariate associations as potential ways to detect data fabrication. The Newcomb-Benford law can be applied when untruncated ratio- or count scale measures are present (**???**); terminal digit analysis can be applied whenever there are sufficient digits (see also **???**). Multivariate associations can be investigated whenever there are two or more variables available and data on that same relation is available from (assumably) genuine data sources.

Detecting data fabrication in summary statistics

P-value analysis

The distribution of a single- or a set of *p*-values is uniform if the null hypothesis is true; it is right-skewed if the alternative hypothesis is true (Fisher, 1925). The distribution of one *p*-value is the result of the population effect size, the precision of the estimate, and the observed effect size, whose properties carry over to a set of independent *p*-values if those *p*-values are independent. As such, the *p*-value distribution of a set of independent *p*-values is uniform when the null hypothesis is true, or right-skewed when the alternative hypothesis is true.

When assumptions underlying the computation of a *p*-value are violated, *p*-value distributions can take on a variety of shapes. For example, when optional stopping occurs and the null hypothesis is true, *p*-values just below .05 become more frequent (**???**; C. H. Hartgerink, Aert, Nuijten, Wicherts, & Assen, 2016). However, when optional stopping occurs under the alternative hypothesis or when other researcher degrees of freedom are used, a right-skewed distribution for significant *p*-values can still occur (**???**, **???**).

When independent p-values are not right-skewed or uniformly distributed (as would be theoretically expected), it can indicate potential data fabrication. For example, in the Fuji case, supposedly randomly assigned groups were fabricated. In truly randomly assigned groups, the measurements are statistically identical (prior to an intervention). However, in the Fuji case Carlisle observed many large p-values, which ultimately led to the identification of potential data fabrication [@]. In Table xxxx we illustrate the difference between expected data under the null distribution (Set 1) and excessively consistent and potentially fabricated data (Set 2). More specifically, the expected value of a uniform p-value distribution is .5, but the fabricated data from our illustration have a mean p-value of 0.956.

Table 1: Examples of means and standard deviations for a continuous outcome in genuine- and fabricated randomized clinical trials. Set 1 (S1) is randomly generated data under the null hypothesis of random assignment (assumed to be the genuine process), whereas Set 2 (S2) is generated under excessive consistency with equal groups. Each trial condition contains 100 participants. The p-values are the result of independent t-tests comparing the experimental and control conditions within each respective set.

Study	M_E (SD_E) [S1]	M_C (SD_C) [S1]	P-value [S1]	M_E (SD_E) [S2]	M_C (SD_C) [S2]	P-value [S2]
Study 1	48.432 (10.044)	49.158 (9.138)	0.594	52.274 (10.475)	63.872 (10.684)	0.918
Study 2	50.412 (10.322)	49.925 (9.777)	0.732	62.446 (10.454)	60.899 (10.398)	0.989
Study 3	51.546 (9.602)	51.336 (9.479)	0.877	62.185 (10.239)	55.655 (10.457)	0.951
Study 4	49.919 (10.503)	50.857 (9.513)	0.509	62.468 (10.06)	68.469 (10.761)	0.956
Study 5	49.782 (11.167)	50.308 (8.989)	0.714	67.218 (10.328)	55.846 (10.272)	0.915
Study 6	48.631 (9.289)	49.29 (10.003)	0.630	62.806 (11.216)	66.746 (11.14)	0.975
Study 7	49.121 (9.191)	47.756 (10.095)	0.318	50.19 (10.789)	55.724 (10.302)	0.960
Study 8	49.992 (9.849)	51.651 (10.425)	0.249	54.651 (11.372)	55.336 (10.388)	0.995
Study 9	50.181 (9.236)	51.292 (10.756)	0.434	63.322 (11.247)	53.734 (11.488)	0.941
Study 10	49.323 (10.414)	49.879 (9.577)	0.695	60.285 (10.069)	54.645 (11.211)	0.960

In order to test whether a distribution of independent p-values might be fabricated, we previously proposed using the Fisher method (Fisher, 1925; S. P. O'Brien et al., 2016). The Fisher method originally was intended as a meta-analytic tool, which tests whether there is sufficient evidence for an effect (i.e., right-skewed p-value distribution). This test is computed as

$$\chi^2_{2k} = -2\sum_{i=1}^{k} \ln(p_i)$$

where it tests for more smaller p-values than larger p-values across the k number of p-values. Reversing this results in

$$\chi^2_{2k} = -2\sum_{i=1}^{k} \ln(1 - \frac{p_i - t}{1 - t})$$

where it now tests for more larger p-values than smaller p-values across the k number of p-values that fall above the threshold t (i.e., the Fisher method now tests for left-skew). When $t = 0$, all p-values are selected. When $t > 0$ the remaining p-values are rescaled to fit the original 0-1 range (i.e., dividing by $1 - t$). This test is similar (but not equivalent) to Carlisle's method testing for excessive homogeneity across baseline measurements in RCTs (Carlisle, 2012, 2017; Carlisle et al., 2015).

As an example, we apply the reversed Fisher method to both the genuine- and fabricated results. Using the threshold $t = 0.05$ to only select the nonsignificant results from Table xxxx, we retain $k = 10$ genuine p-values and $k = 10$ fabricated p-values. This results in $\chi^2_{2\times10} = 18.362, p = 0.564$ for the genuine data from Table xxxx; $\chi^2_{2\times10} = 66.848, p = 0$ for the fabricated data from Table xxxx. Another more practical example directly from the Fuji case (Carlisle, 2012), anecdotally illustrates that actual fabricated data can

result in significant findings with the reversed Fisher method. For example, p-values extracted from the original Table 3 (fentanyl dose; Carlisle, 2012) for five independent comparisons show excessively high p-values, $\chi^2_{2\times5} = 19.335, p = 0.036$.

We note that misspecified one-tailed tests can also result in excessive amounts of large p-values. For correctly specified one-tailed tests, the p-value distribution is right-skewed if the alternative hypothesis is true. When the alternative hypothesis is true, but the effect is in the opposite direction of the hypothesized effect (e.g., a negative effect when a one-tailed test for a positive effect is conducted), this results in a left-skewed p-value distribution. As such, any data fabrication detected with this method would need to be inspected for misspecified one-tailed hypotheses to preclude false conclusions. In the studies we present in this manuscript, misspecification of one-tailed hypothesis testing is not an issue because we prespecified the effect and its direction to the participants.

Variance analysis

Sample variance- or standard deviation estimates are typically reported to indicate dispersion, but just like the mean there should be sampling error in this estimate proportional to the sample size (i.e., $\sigma/\sqrt{2n}$ under the assumption of normality, p. 351, Yule, 1922). If an observed random variable x is normally distributed, the variance of x is χ^2-distributed (p. 445; Hogg & Tanis, 2001); that is

$$var(x) \sim \frac{\chi^2_{N_j-1}}{N_j - 1}$$

where N is the sample size of the jth group. Using the reported summary statistics, we can compute the Mean Squares within (MS_w) each condition as

$$MS_w = \frac{\sum_{j=1}^{k}(N_j - 1)s_j^2}{\sum_{j=1}^{k}(N_j - 1)}$$

where s_j^2 is the variance in the jth group. The reported variances can be standardized by dividing the observed variances by the MS_w; we denote standardized variances with z^2.

By standardizing the reported variances observed dispersion across measures can be calculated. Observed dispersion of the standardized variances can be operationalized as the standard deviation of the variances (denoted in this paper as SD_z, Simonsohn, 2013) or as the range of the variances (denoted as $max - min_z$). Note that such comparisons are only meaningful if the reported variances originate from a homogeneous population distribution of variances (otherwise, subgroups need to be made).

In order to compute the expected dispersion, we use the distribution of the standardized variances to compute how extreme the observed dispersion of the variances is. Assuming normality, the distribution of the standardized variances follows a χ^2-distribution in the form of

$$z_j^2 \sim \left(\frac{\chi^2_{N_j-1}}{N_j - 1}\right) / MS_w$$

which is the result of the distribution of $var(x)$ as denoted earlier, divided by the MS_w. From this distribution, fictitious variances can be generated for each group under investigation. These are subsequently used to compute one of the measures for dispersion proposed in the previous paragraph (e.g., standard deviation of the variances). Repeating this across i iterations provides an estimate of the density function for the expected dispersion of the variances. By comparing the observed dispersion of the variances with the expected variances, we can estimate how extreme our observations are. For our purposes, too little dispersion in the variances indicates potential fabrication in the reported data (Simonsohn, 2013).

As an example, we apply the variance analysis to the illustration from Table xxxx and the Smeesters case. For the reported standard deviations in Table xxxx, we apply the variance analysis across those in the genuine- and fabricated sets separately. For the genuine data (Set 1), we find that the reported mean standard deviation is 9.8683942 with a standard deviation of 0.5952913; for the fabricated data (Set 2), we find that the reported mean standard deviation is 10.6674101 with a standard deviation of 0.4555638. These summary statistics of the summary statistics already indicate there is a difference between the genuine- and fabricated data. Variance analysis helps us quantify this: Set 1 has no excessive consistency (0.217), whereas Set 2 does show excessive consistency (0.006). In words, out of 10^{4} theoretically expected samples, 2174 showed more consistency for Set 1, whereas only 56 showed more consistency for Set 2. As a non-fictional example, three independent conditions from the same study ($n_k = 15$) were reported to have standard deviations 25.09, 24.58, and 25.65 in the Smeesters case. The standard deviation of these standard deviations is 0.54 (i.e., SD_z). Such consistency (or even more) would only be observed in 1.37% of 100,000 simulated replications (Simonsohn, 2013).

Study 1 - detecting fabricated summary statistics

We tested the performance of statistical methods to detect data fabrication in summary statistics with genuine- and fabricated summary statistics from four anchoring studies (Jacowitz & Kahneman, 1995; A. Tversky & Kahneman, 1974). The anchoring effect is a well-known psychological heuristic that uses the information in the question as the starting point for the answer, which is then adjusted to yield a final estimate of a quantity. For example 'Do you think the percentage of African countries in the UN is above or below [10% or 65%]? What do you think is the percentage of African countries in the UN?'. These questions yield mean responses of 25% and 45%, respectively (A. Tversky & Kahneman, 1974), despite essentially posing the same factual question. A considerable amount of (assumably) genuine data sets on the anchoring heuristic are freely available (https://osf.io/pqf9r; Klein et al., 2014) and we collected fabricated data sets within this study. This study was approved by the Tilburg Ethical Review Board (EC-2015.50).

Methods

We collected summary statistics for four anchoring studies: (i) distance from San Francisco to New York, (ii) population of Chicago, (iii) height of the Mount Everest, and (iv) the number of babies born per day in the United States (Jacowitz & Kahneman, 1995). Each of the four studies provided us with summary statistics for a 2 (low/high anchoring) $\times$ 2 (male/female) factorial design. Throughout our study, the unit of analysis is a set of summary statistics (i.e., means, standard deviations, and test results) for the four anchoring studies from one respondent. For current purposes, a respondent is defined as researcher/lab where the four anchoring studies' summary statistics originate from. All materials, data, and analyses scripts are freely available on the OSF (https://osf.io/b24pq) and a preregistration is available at https://osf.io/tshx8/. Throughout this report, we will indicate which facets were not preregistered or deviate from the preregistration by denoting "(not preregistered)" or "(deviation from preregistration)", respectively.

Data collection

We downloaded thirty-six genuine data sets from the publicly available Many Labs (ML) project (https://osf.io/pqf9r; Klein et al., 2014). The ML project replicated several effects across thirty-six locations, including the anchoring effect in the four studies mentioned previously. Considering the size of the ML project, the transparency of research results, and minimal individual gain for fabricating data, we assumed these data to be genuine. For each of the thirty-six locations we computed three summary statistics (i.e., sample sizes, means, and standard deviations) for each of the four conditions in the four anchoring studies (i.e., $3 \times 4 \times 4$) for each of the thirty-six locations. We computed these summary statistics from the raw ML data, which were cleaned using the original analysis scripts from the ML project.

The sampling frame consisted of 2,038 psychology researchers who published a peer-reviewed paper in 2015, as indexed in Web of Science (WoS) with the filter set to the U.S. We sampled psychology researchers to improve familiarity with the anchoring effect (Jacowitz & Kahneman, 1995; A. Tversky & Kahneman, 1974), for which summary statistics were fabricated. We filtered for U.S. researchers to ensure familiarity with the imperial measurement system, which is the scale of some of the anchoring studies and in order to reduce heterogeneity across fabricators. We searched WoS on October 13, 2015. In total, 2,038 unique corresponding e-mails were extracted from 2,014 papers (due to multiple corresponding authors).

References

Bik, E. M., Casadevall, A., & Fang, F. C. (2016). The prevalence of inappropriate image duplication in biomedical research publications. *mBio*, *7*(3), e00809–16. http://doi.org/10.1128/mbio.00809-16

Carlisle, J. B. (2012). The analysis of 168 randomised controlled trials to test data integrity. *Anaesthesia*, *67*(5), 521–537. http://doi.org/10.1111/j.1365-2044.2012.07128.x

Carlisle, J. B. (2017). Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. *Anaesthesia*. http://doi.org/10.1111/anae.13938

Carlisle, J. B., Dexter, F., Pandit, J. J., Shafer, S. L., & 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. http://doi.org/10.1111/anae.13126

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, *4*(5), e5738. http://doi.org/10.1371/journal.pone.0005738

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburg, United Kingdom: Oliver Boyd.

Haldane, J. B. S. (1948). The faking of genetical results. *Eureka*, *6*, 21–28. Retrieved from http://wayback.archive.org/web/20170206144438/http://www.archim.org.uk/eureka/27/faking.html

Hartgerink, C. H., Aert, R. C. van, Nuijten, M. B., Wicherts, J. M., & Assen, M. A. van. (2016). Distributions of p-values smaller than .05 in psychology: What is going on? *PeerJ*, *4*, e1935. http://doi.org/10.7717/peerj.1935

Hobbes, T. (1651). *Leviathan*. Oxford University Presss.

Hogg, R. V., & Tanis, E. A. (2001). *Probability and statistical inference*. New Jersey, NJ: Prentice-Hall.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality & Social Psychology Bulletin*, *21*, 1161–1166. http://doi.org/10.1037/e722982011-058

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Bernstein, M. J., ... Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, *45*(3), 142–152. http://doi.org/10.1027/1864-9335/a000178

Koppers, L., Wormer, H., & 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*. http://doi.org/10.1007/s11948-016-9841-7

Levelt. (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel*. Retrieved from https://www.commissielevelt.nl/

Oransky, I. (2015). The Retraction Watch Leaderboard. Retrieved from http://wayback.archive.org/web/20170206163805/http://retractionwatch.com/the-retraction-watch-leaderboard/

O'Brien, S. P., Danny Chan, Leung, F., Ko, E. J., Kwak, J. S., Gwon, T., ... Bouter, L. (2016). Proceedings of the 4th world conference on research integrity. *Research Integrity and Peer Review*, *1*(S1). http://doi.org/

10.1186/s41073-016-0012-9

Parker, A., & Hamblen, J. (1989). Computer algorithms for plagiarism detection. *IEEE Transactions on Education*, *32*(2), 94–99. http://doi.org/10.1109/13.28038

Simonsohn, U. (2013). Just post it. *Psychological Science*, *24*(10), 1875–1888. http://doi.org/10.1177/0956797613480366

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

The Journal of Cell Biology. (2015). About the Journal. Retrieved from https://web.archive.org/web/20150911132421/http://jcb.rupress.org/site/misc/about.xhtml

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. http://doi.org/10.1037/h0031322

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. http://doi.org/10.1126/science.185.4157.1124

Yule, G. U. (1922). An introduction to the theory of statistics. Retrieved from https://ia800205.us.archive.org/13/items/cu31924013993187/cu31924013993187.pdf