How many researchers use 'boilerplate' statistical analysis sections?

An observational study of papers published in *PLOS ONE*.

Nicole White, Thiru Balasubramaniam, Richi Nayak, Adrian Barnett

21/05/2020

An ideal statistical analysis will use appropriate methods to create insights from the data and inform the research questions. Unfortunately many current statistical analyses are far from ideal, with many researchers using the wrong methods, misinterpreting the results, or failing to adequately check their assumptions (Goodman 2008; Leek et al. 2017). Some researchers take a "mechanistic" approach to statistics, copying the few methods they know regardless of their appropriateness, and then going through the motions of the analysis (Stark and Saltelli 2018).

Many researchers lack adequate training in research methods, and statistics is something they do with trepidation and even ignorance (Altman 1994; King et al. 2019). However, using the wrong statistical methods can cause real harm (Altman 1994) and bad statistical practices are being to used abet weak science (Stark and Saltelli 2018). Statistical mistakes are a key source of waste in research and partly explain the current reproducibility crisis in science (Allison et al. 2016). Even when the correct methods are used, many researchers fail to describe them adequately, making it difficult to reproduce the results (Ernst and Albers 2017; Zhou and Skidmore 2018).

The International Committee of Medical Journal Editors recommend that researchers should: "Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to judge its appropriateness for the study and to verify the reported results" (ICJME 2019). Although the general lack of statistical understanding from both authors and reviewers means this recommendation may not be checked. A recent survey of editors found that only 23% of health and medical journals used expert statistical review for all articles (Hardwicke and Goodman 2020), which was little different from a survey from 22 years ago (Goodman, Altman, and George 1998).

Two statisticians on this paper (AB and NW) have heard researchers admit that they sometimes copy-and-paste their statistical methods sections from other papers, regardless of whether they are appropriate. The aim of this paper is to use text-mining methods to estimate the extent that researchers are using cut-and-paste or 'boilerplate' statistical methods sections. Use of these methods sections indicates that little thought has gone into the statistical analysis.

1 Methods

1.1 Data sources

We used two openly available data sources to find statistical methods sections.

1.1.1 Public Library of Science (PLOS ONE)

PLOS ONE is a large open access journal that publishes original research across a wide range of scientific fields. Articles must be in English. Article submissions are handled by an academic editor who selects peer

reviewers based on their self-nominated area(s) of expertise. Submissions do not undergo formal statistical review. Instead, reviewers are required to assess submissions against several publication criteria, including whether: "Experiments, statistics, and other analyses are performed to a high technical standard and are described in sufficient detail". Authors are encouraged to follow published reporting guidelines to ensure that chosen statistical methods are appropriate for the study design, and adequate details are provided to enable independent replication of results. Reporting guidelines are available from the EQUATOR network which has developed guidelines for study designs commonly used in health research, and these guidelines were developed to tackle poor statistical application and reporting in health and medical journals (Altman and Simera 2016).

All *PLOS ONE* articles are freely accessible via the PLOS Application Programming Interface (API). This functionality enabled us to conduct semi-automated searches of full-text articles and analyse data on individual records, including text content and general attributes such as publication date and field(s) of research. To find papers with a statistical methods section we used targeted API searches followed by article filtering based on section headings. The data were downloaded on 3 July 2020.

Step 1: Targeted API searches. API searches were completed using the R package 'rplos' (Chamberlain, Boettiger, and Ram 2020). Search queries targeted the presence of analysis-related terms anywhere in the article. Search terms combined the words "data" or "statistical" with one of: "analysis", "analyses", "method", "methodology" or "model(l)ing". Search terms were intended to be broad whilst keeping search results to a manageable number for full-text review (see Step 2). By allowing terms to appear anywhere within the article, we accounted for the possibility of relevant text being placed in different sections, for example, in the *Material and Methods* section versus *Results*. Search results were indexed by a unique Digital Object Identifier (DOI). Attribute data collected per DOI included journal volume, subject classification(s) and total article views since publication.

Step 2: Partial matching on section headings. Full text XML data for all search results were downloaded and combined into a single dataset, organised by DOI and subsection heading(s). Since *PLOS ONE* does not prescribe standardised headings to preface statistical methods sections, we performed partial matching on available headings against frequently used terms in initial search results: 'Statistical analysis', 'Statistical analyses', 'Statistical method', 'Statistics', 'Data analysis' and 'Data analyses'. To determine the reliability of our chosen filters, we manually reviewed full text data extracted for a random sample of XXX articles that were not matched (File S1).[TODO... finish this thought...]

1.1.2 Australia and New Zealand Clinical Trials Registry (ANZCTR)

The ANZCTR was established in 2005 as part of a coordinated global effort to improve research quality and transparency in clinical trials reporting; observational studies can also be registered. All studies registered on ANZCTR are publicly available and can be searched via an online portal (https://www.anzctr.org.au/). Details required for registration follow a standardised template (reference or supp file), which covers participant eligibility, the intervention(s) being evaluated, study design and outcomes. Researchers are asked to provide a "brief description" of the sample size calculations, statistical methods and planned analyses, although this section is not compulsory (anzctr 2019). The information provided must be in English. Studies are reviewed by ANZCTR staff for completeness of key information, which does not include the completeness of the statistical methods sections. Studies are not peer reviewed.

All studies available on ANZCTR were downloaded on 1 February 2020 in XML format. We used all the text available in the "Statistical methods" section. We also collated basic information about the study including the study type (interventional or observational), submission date, number of funders and target sample size. These variables were chosen as we believed they might influence the completeness of the statistical methods section, because we expected larger studies and those with funding to be more complete, and we also were interested in changes over time. Studies prior to 2013 were excluded as the statistical methods section appeared to be introduced in 2013. Some studies were first registered on the alternative trial database *clinicaltrials.gov* and then also posted to ANZCTR. We excluded these studies because they almost all had no completed statistical methods section as this requested by *clinicaltrials.gov*.

1.2 Full-text processing

Text cleaning aimed to standardise notation and statistical terminology, whilst minimising changes to article style and formatting. Full details of text cleaning steps undertaken and corresponding R code are provided in Supplementary File X.

Mathematical notation including Greek letters was converted from Unicode characters to plain text. For example, the Unicode characters corresponding to θ (<U+03B8>) was replaced with 'theta'. Similarly, common symbols outside of Unicode blocks including '%' (percent) and '<' ('less-than') were converted into plain text, using functions available in the 'textclean' package (Rinker 2018). General formatting was removed, this included carriage returns, punctuation marks, in-text references (e.g. "[42]") centred equations, and other non-ascii characters. Text contained inside brackets was retained in the dataset to maximise content for analysis, with brackets removed.

We compiled an extensive list of statistical terms to standardise descriptions of statistical methods reported across both datasets. An initial list was compiled by calculating individual word frequencies and identifying relevant terms that appeared at least 100 times. Further terms were sourced from index searches of statistics reference textbooks [ref]. The final list is provided as Supplementary Material. Possible variants including plurals (e.g. 'chi-squares') unhyphenated (e.g 'chi square') and combined (e.g. 'chisquare') terms were transformed to singular, hyphenated form (e.g 'chi-square'). Common statistical tests were also hyphenated (e.g. 'hosmer lemeshow' to 'hosmer-lemeshow').

As a final step, common stop words including pronouns, contractions and selected prepositions were removed. We retained selected stop words that, if excluded, may have changed the context of statistical method(s) being described, for example 'between' and 'against'.

1.3 Clustering algorithm

1.4 Missing statistical methods sections

The statistical methods section for the ANZCTR data was somtimes missing and so we examined if there were particular studies where this section was more likely to be missing. We used a logistic regression model fitted using a Bayesian paradigm. A small number of sections were labelled as "Not applicable", "Nil" or "None" and we changed these to missing.

2 Results

2.1 PLOS ONE

API searches returned 131847 unique records, of which 111731 (85%) included a statistical methods section based on our search criteria.

Search results varied by journal volume (Figure 1A). The total number of API search results peaked at volumes 8 (n = 19045) and 9 (n = 19045), corresponding to years 2013 and 2014. This trend aligned with the total number of papers published in PLOS ONE over the same timeframe. The percentage of records that included a statistical methods section by volume based on our proposed matching criteria varied between 64% (volume 2) and 86% (volume 9). [TODO: breakdown of number of records that had stats methods in supp material, other observations about those excluded at second stage filter]

The median length of statistical methods sections was 127 words (IQR: 61 to 254 words) (Figure 1B). 7450 articles (7%) had a statistical methods section of 500 words or more. 19461 articles (17%) had a statistical methods section of 50 words or less, equal to the length of this paragraph.

All papers included Biology and life sciences (n = 107584), Earth sciences (n = 7605) and/or Computer and information sciences (n = 5190) within their top 3 subject classifications (Figure 1C).

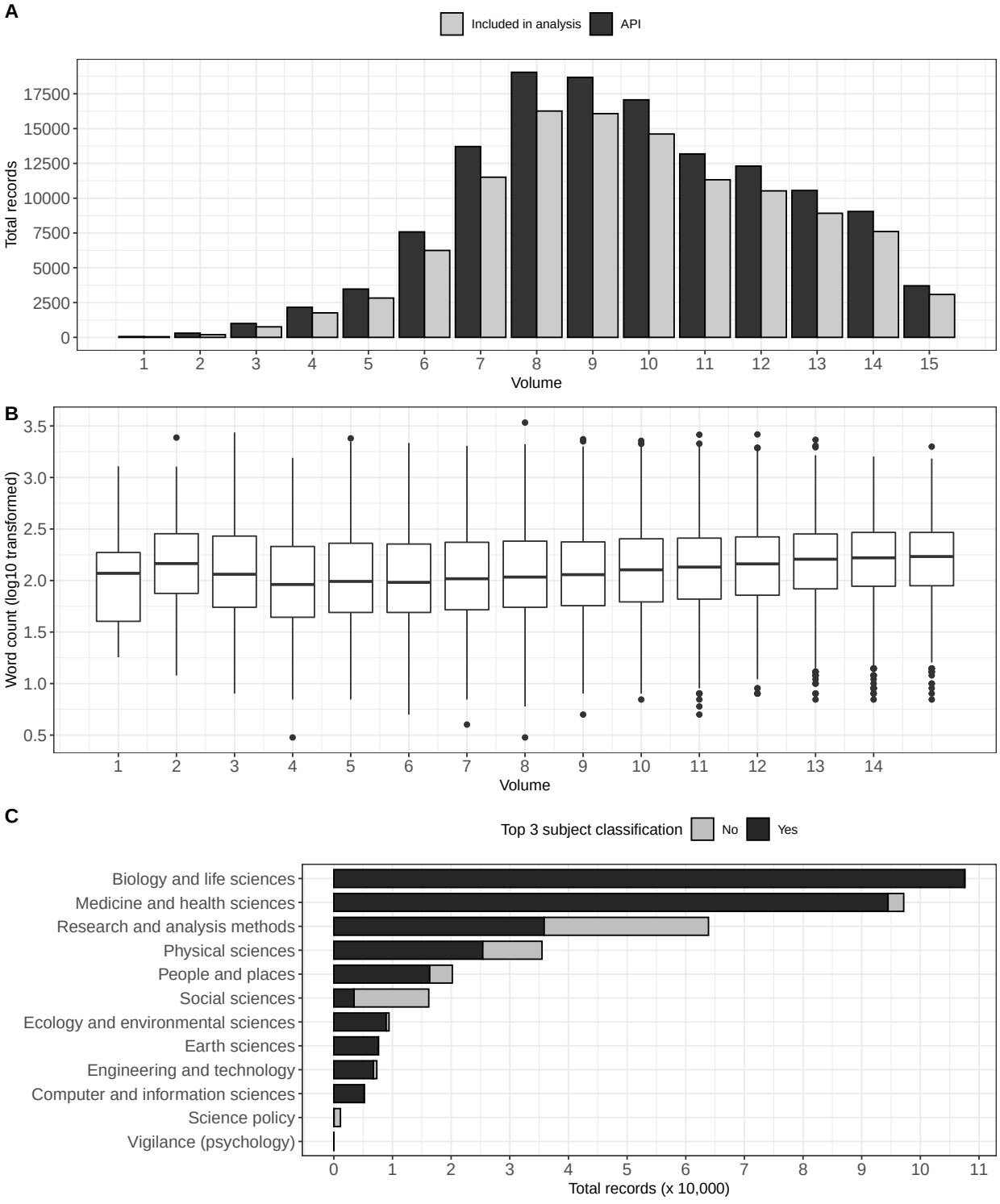

Figure 1: Search results by PLOS ONE volume (1st row); word count per statistical methods section included in analysis (n = 111,731; 2nd row); subject classifications assigned to full-text records included in analysis (3rd row)

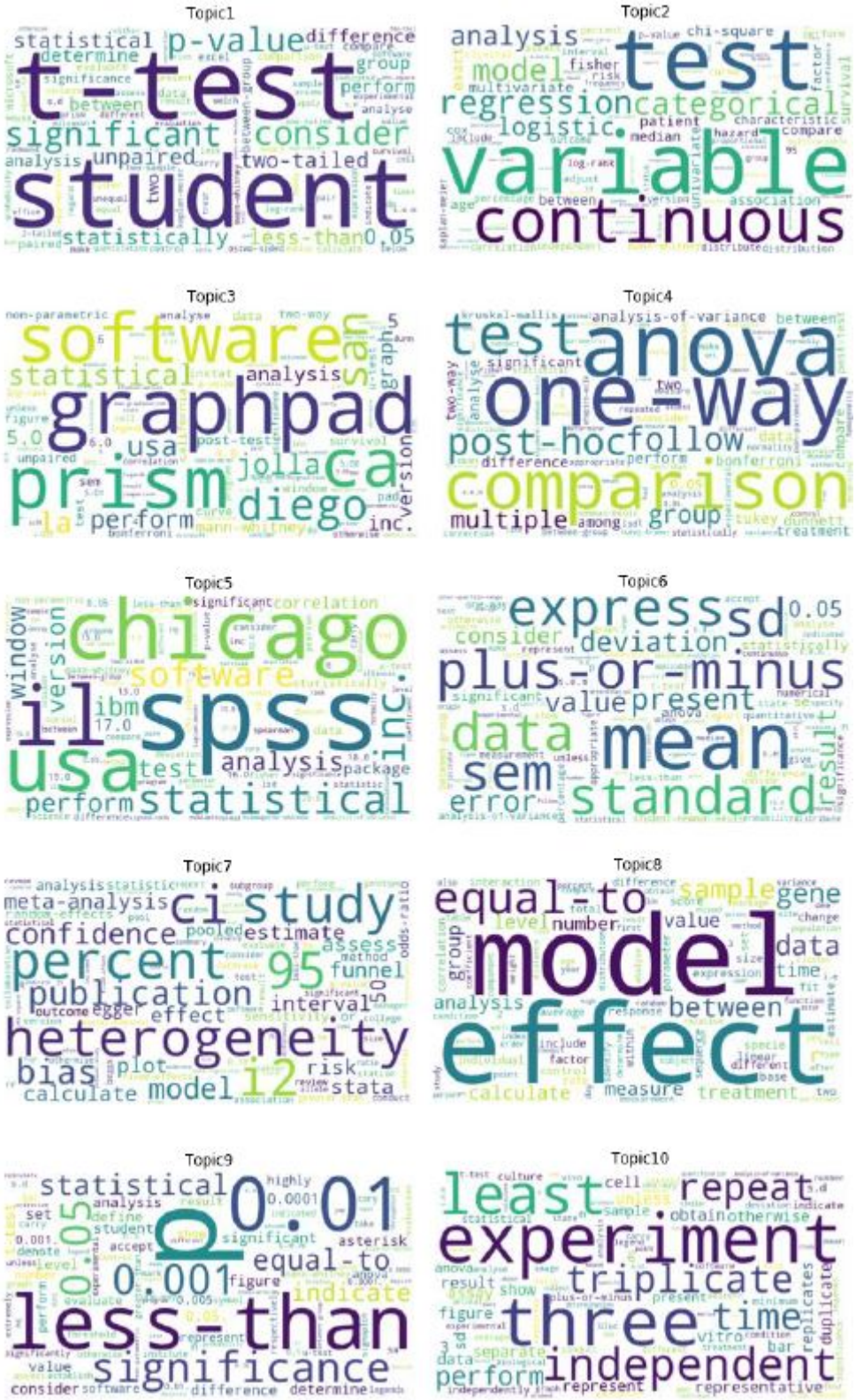

Figure 2: Topic clouds summarising common terms in PLOS ONE statistical methods sections assuming ten clusters

We applied the clustering algorithm to the cleaned dataset, varying the number of clusters from 1 to 50. Increasing the number of clusters decreased cluster quality based on global goodness-of-fit measures (Supplementary Figure 1), with average silhouette score and within-cluster dispersion leveling off around 20 clusters. This indicated that the data comprised one large, heterogeneous cluster and multiple smaller clusters.

Figure 2 displays topic clouds based on ten clusters. General themes reflected the use of statistical software, between-group hypothesis testing, definitions of statistical significance and regression modelling.

Two topics reflected the use of statistical software: Prism GraphPad (Topic 3: n = 9,879; 8.8%) and SPSS (Topic 5: n = 9,574; 8.6%) (Box 1). Manual review of top matching sections assigned to Topic 3 showed evidence of boilerplate text. Out of the top 10 matches, 9 did not specify which statistical methods were used.

Box 1: Examples of boilerplate text (Statistical software)

Topic 3

- *graphpad prism (graphpad software, san diego, ca) was used for all analyses.*
- *all statistical analysis was performed using the graphpad prism software.*
- *statistical analysis were done using graphpad prism software (graphpad software, inc., la jolla, ca).*

Topic 5

- *all statistical analyses were conducted in spss 19 (spss inc., chicago, il, usa).*
- *all statistical analyses were performed using the spss version 15 for windows (spss inc., chicago, il, usa). a p-value less than 0.05 was considered a significant difference.*
- *statistical analysis was performed using the spss 17.0 software package (spss, chicago, il, usa). all data were analyzed using t tests, values of $p < 0.05$ were considered significant.*

Evidence of boilerplate text was identified among topics that focused on two-group hypothesis testing as the only statistical method used (Box 2). Topic 1 reflected applications of Student's t-test assuming a 5% level of statistical significance (n = 3,775). Similar methods sections were identified under Topic 9 (n = 6,104) but were differentiated by multiple thresholds for declaring statistical significance: *p<0.05, **p<0.01 and ***p<0.001.

Box 2: Examples of boilerplate text (Between-group hypothesis testing, two groups)

Topic 1

- *student's t-test was used for statistical analysis. a p value of <0.05 was considered statistically significant.*
- *statistical analysis was performed using student's t-test. a p-value <0.05 was considered to be statistically significant.*
- *statistical analysis was performed using a two-tailed student's t-test, and p-values <0.05 were considered statistically significant.*

Topic 9

- *statistical significance was determined by students t-tests. *p<0.05, ** p = <0.01, *** p<0.001.*
- *statistical analyses were performed using the nonparametric mann-whitney test. differences were considered significant for *p<0.05, ** p<0.01, ***p<0.001.*
- *comparisons between two groups were performed by student's test. statistical significance was defined as *p<0.05;* ** p<0.01; *** p<0.001.*

Expanded descriptions of between-group analysis were found in three topics (4, 6 and 10), varying based on method(s) used and study design (Box 3 - include?). Sections assigned to these topics appeared to follow the general template of reporting descriptive statistics followed by methods of hypothesis testing depending on the number of groups being compared. Sections strongly associated with Topic 6 (n = 4,746) combined descriptive statistics with follow-up testing of group differences by Student's t-test, Mann Whitney U or one-way analysis of variance. Descriptive statistics were commonly reported as means with standard deviations or standard errors. Topic 10 (n = 4,664) reported similar methods applied to the analysis of replicates obtained under experimental designs. Variations to descriptions of one-way analysis of variance were identified under Topic 4 (n = 10,163), with an emphasis on post-hoc multiple comparisons.

Remaining topics reflected applications of meta-analysis (Topic 7: n = 5,165), regression modelling (Topic 2: n = 19,478) and larger sections that reported data processing and statistical methods within the same section (Topic 8: n = 38,183). Top matches for Topic 2 shared features of the 'describe-compare' template seen in other topics, expanded to include details of multivariable regression modelling. Examples of multivariable models were (stepwise) logistic regression and cox proportional hazards regression.

2.2 ANZCTR

We downloaded 28,008 studies. The numbers of excluded studies are shown in Figure~X. Of the 12,700 included studies, 9,523 had a statistic section which is 75% (95% CI 74% to 76%).

We examined if four study characteristics were associated with a missing statistics section. The odds ratios and 95% credible intervals are in Table~X. Observational studies were less likely to have a missing methods section compared with interventional studies. Missing sections became less likely over time. Studies with more funders and a larger target sample size were less likely to have a missing methods section.

Variable & Odds ratio & 95% CI Study type = Observational & 0.78 & 0.69, 0.89 Date (per year) & 0.90 & 0.88, 0.91 Number of funders & 0.80 & 0.74, 0.86 Target sample size (per doubling) & 0.90 & 0.88, 0.92

Some of the non-missing statistics sections were only one word, including "ANOVA", "t-test", "SPSS" and even "SSPS". The median length of the section was 129 words with an inter-quartile range of 71 to 219 words.

- Final sample size

3 Discussion

Many researchers are using lazy practice by copying a standard "boilerplate" statistical methods section, likely cut-and-pasting from other researchers or projects. This is a strong sign of the ritualistic practice of statistics where researchers go through the motions rather than using conscientious practice (Stark and Saltelli 2018). This is concerning because using the wrong statistical methods can reduce the potential of study, or worse, invalidate the entire study. Poor statistical practice is a key driver of the ongoing reproducibility crisis in science (Ioannidis et al. 2014).

The first line in many statistical analysis sections in *PLOS ONE* was the software used and some entire sections in ANZCTR only stated the software, implying that the software is the most important detail. As

Doug Altman said, "Many people think that all you need to do statistics is a computer and appropriate software" (Altman 1994). This is not the case, and whilst it is important for researchers to mention the software and version used for reproducibility purposes, it is a minor detail compared with detailing what methods were used and why.

Despite the extensive array of statistical tests available, many authors are reporting the same few methods.

One reason these inadequate sections get published is that most journals do not use statistical reviewers, despite empirical evidence showing they improve manuscript quality (Hardwicke and Goodman 2020).

3.1 Limitations

We did not check whether papers used the correct methods, and for some simple studies a 'boilerplate' statistical methods section would be fine.

We examined papers where there was a statistics section, and we missed papers that used statistical analysis but did not include a statistical analysis section. Reiterate outcomes of random sample checking here.

We only examined one journal and one trial registry and hence our results may not be generalisable to all journals or registries, especially those that consistently use a statistical reviewer.

We searched the full text of *PLOS ONE* papers but not the supporting information which may contain statistical methods sections for some papers. The search terms we used to find statistical methods appeared in the supporting information titles for xxx papers (x%). We did not include the supporting information because it is less structured than the paper and could be in PDF or Word format.

References

Allison, David B., Andrew W. Brown, Brandon J. George, and Kathryn A. Kaiser. 2016. "Reproducibility: A Tragedy of Errors." *Nature* 530 (7588): 27–29. https://doi.org/10.1038/530027a.

Altman, D G. 1994. "The Scandal of Poor Medical Research." *BMJ* 308 (6924): 283–84. https://doi.org/10.1136/bmj.308.6924.283.

Altman, Douglas G, and Iveta Simera. 2016. "A History of the Evolution of Guidelines for Reporting Medical Research: The Long Road to the EQUATOR Network." *Journal of the Royal Society of Medicine* 109 (2): 67–77. https://doi.org/10.1177/0141076815625599.

anzctr. 2019. "ANZCTR Data Field Definitions V25." https://www.anzctr.org.au/docs/ANZCTR%20Data%20field%20explanation.pdf.

Chamberlain, Scott, Carl Boettiger, and Karthik Ram. 2020. *Rplos: Interface to the Search API for 'PLoS' Journals*. https://CRAN.R-project.org/package=rplos.

Ernst, Anja F., and Casper J. Albers. 2017. "Regression Assumptions in Clinical Psychology Research Practicea Systematic Review of Common Misconceptions." *PeerJ* 5: e3323. https://doi.org/10.7717/peerj.3323.

Goodman, Steven. 2008. "A Dirty Dozen: Twelve P-Value Misconceptions." *Seminars in Hematology* 45 (3): 135–40. https://doi.org/10.1053/j.seminhematol.2008.04.003.

Goodman, Steven N., Douglas G. Altman, and Stephen L. George. 1998. "Statistical Reviewing Policies of Medical Journals." *Journal of General Internal Medicine* 13 (11): 753–56. https://doi.org/10.1046/j.1525-1497.1998.00227.x.

Hardwicke, Tom E, and Steve Goodman. 2020. "How Often Do Leading Biomedical Journalsuse Statistical Experts to Evaluate Statistical Methods? The Results of a Survey." MetaArXiv. https://doi.org/10.31222/osf.io/z27u4.

ICJME. 2019. "Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals." http://www.icmje.org/icmje-recommendations.pdf.

Ioannidis, John P A, Sander Greenland, Mark A Hlatky, Muin J Khoury, Malcolm R Macleod, David Moher, Kenneth F Schulz, and Robert Tibshirani. 2014. "Increasing Value and Reducing Waste in Research Design, Conduct, and Analysis." *The Lancet* 383 (9912): 166–75. https://doi.org/10.1016/s0140-6736(13)62227-8.

King, Kevin M., Michael D. Pullmann, Aaron R. Lyon, Shannon Dorsey, and Cara C. Lewis. 2019. "Using Implementation Science to Close the Gap Between the Optimal and Typical Practice of Quantitative Methods in Clinical Science." *Journal of Abnormal Psychology* 128 (6): 547–62. https://doi.org/10.1037/abn0000417.

Leek, Jeff, Blakeley B. McShane, Andrew Gelman, David Colquhoun, Michèle B. Nuijten, and Steven N. Goodman. 2017. "Five Ways to Fix Statistics." *Nature* 551 (7682): 557–59. https://doi.org/10.1038/d41586-017-07522-z.

Rinker, Tyler W. 2018. *textclean: Text Cleaning Tools*. Buffalo, New York. https://github.com/trinker/textclean.

Stark, Philip B., and Andrea Saltelli. 2018. "Cargo-Cult Statistics and Scientific Crisis." *Significance* 15 (4): 40–43. https://doi.org/10.1111/j.1740-9713.2018.01174.x.

Zhou, Yuanyuan, and Susan Skidmore. 2018. "A Reassessment of ANOVA Reporting Practices: A Review of Three APA Journals." *Journal of Methods and Measurement in the Social Sciences* 8 (1): 3–19. https://doi.org/10.2458/v8i1.22019.