An observational study of papers published in *PLOS ONE* and studies posted to a trial registry.

Nicole White *
School of Public Health and Social Work, QUT
and
Thiru Balasubramaniam, Richi Nayak
To add
and
Adrian Barnett
School of Public Health and Social Work, QUT

May 13, 2021

Abstract

The text of your abstract. 200 or fewer words.

Keywords: 3 to 6 keywords, that do not appear in the title

*The authors gratefully acknowledge computational resources and services used in this work provided by the eResearch Office, Queensland University of Technology, Brisbane, Australia.

1 Introduction

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 (?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 & 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, Brown et al. 2018) and bad statistical practices are being to used abet weak science (Stark & 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 & Albers 2017, Zhou & Skidmore 2018). Poor statistical methods might not be caught by reviewers, as they may not be qualified to judge the statistics. A recent survey of editors found that only 23% of health and medical journals used expert statistical review for all articles (Hardwicke & Goodman 2020), which was little different from a survey from 22 years ago (Goodman et al. 1998).

There is guidance for researchers on how to write up their statistical methods and results. 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). More detailed guideance is given by the SAMPL and EQUATOR guidelines (Lang & Altman 2013, Altman & Simera 2016) with the latter covering all apsects of the paper. Both of these guidelines were led by Doug Altman, who spoke often and for many years about the need for better statistical reporting. The awareness and use of these guidelines could be improved. There were 256 Google Scholar citations to the SAMPL paper (as at 15 March 2021) which is a good citation statistic for most papers, but is low considering the millions of papers that use statistical analysis.

Two statisticians on this paper (AB and NW) have heard researchers admit that they have copied-and-pasted 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 analyse the content of statistical methods sections included in common research outputs. Results are then evaluated to estimate the extent that researchers are using cut-and-paste or 'boilerplate' statistical methods sections. Boilerplate text is that "which can be reused in new contexts or applications without significant changes to the original" (Wikipedia 2021). Use of these methods sections indicates that little thought has gone into the statistical analysis.

2 Methods

2.1 Data sources

We used two openly available data sources to find statistical methods sections: research articles published in *PLOS ONE* and study protocols registered on the Australian and New Zealand Clinical Trials Registry (ANZCTR). Data sources were chosen as examples of common research outputs that should include descriptions of statistical methods that were used, or are planned, for analysing study outcomes.

2.1.1 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. Article submissions are handled by an academic editor who selects peer reviewers based on their self-nominated areas 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" (PLOS 2021). All reviewers are asked the question: "Has the statistical analysis been performed appropriately and rigorously?", with the possible responses of "Yes", "No" and "I don't know".

Authors are encouraged to follow published reporting guidelines such as EQUATOR, to ensure that chosen statistical methods are appropriate for the study design, and adequate details are provided to enable independent replication of results.

All *PLOS ONE* articles are freely accessible via the PLOS Application Programming Interface (API). This enabled us to conduct searches of full-text articles and analyse data on articles' 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 'rplos' package (Chamberlain et al. 2020). Search queries targeted analysis-related terms, combining the words "data" or "statistical" with one of: "analysis", "analyses", "method", "methodology" or "model(l)ing". We allowed terms to appear anywhere in the article, to allow 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 and subject classification(s).

Step 2: Partial matching on section headings. Full text XML data 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'.

2.1.2 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 (ANZCTR 2019), which covers participant eligibility, the intervention(s) being evaluated, study design and outcomes. The information provided must be in English. Studies are not peer reviewed.

For the statistical methods section, 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). Studies are reviewed by ANZCTR staff for completeness of key information, which does not include the completeness of the statistical methods sections.

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 section is not included in *clinicaltrials.gov*.

2.2 Full-text processing

Text cleaning aimed to standardise notation and statistical terminology, whilst minimising changes to article style and formatting. Datasets were cleaned using the 'textclean package' (Rinker 2018). *R* code used for data extraction and cleaning is available from `https://github.com/agbarnett/stats_section`.

Mathematical notation including Greek letters was converted from Unicode characters to plain text. For example, the Unicode character corresponding to θ (<U+03B8>) was replaced with 'theta'. Common symbols outside of Unicode blocks including '%' (percent) and '<' ('less-than') were similarly converted into plain text. 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 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 three statistics textbooks (Diggle et al. 2013, Bland 2015, Dobson & Barnett 2018). The final list is provided in Supplementary File 1. 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 methods being described, for example 'between' and 'against'.

2.3 Clustering algorithm

Let $P \in R^{M \times N}$ denotes the paper content matrix where the statistical methods sections of N papers consist of M distinct terms. Text clustering for identifying common writing themes in these papers requires this matrix to represent with a vector space model. Let Matrix P be modeled with the unique terms in the processed sections represented with the tf*idf (term frequent * inverse document frequency) weighting schema to consider both common and rare terms.

Text clustering is known to face the curse of dimensionality due to the high number of terms in doc×term matrix representation (**??**). Therefore text-based methods based on distance, density or probability face difficulties (**???**). Specifically, the distance difference between near and far points becomes negligible in high-dimensional data (**?**). This directly affects the distance-based methods such as k-means (**?**) in accurately identifying the common subgroups. In addition, the sparseness of this high dimensional matrix representation does not allow to differentiate the user groups based on density differences (**??**).

Non-negative Matrix factorization (NMF) which maps the high-dimensional data to a lower-dimensional space has been found to provide an effective solution by allowing to form clusters in the lower-dimensional space (**??**). NMF approximates topic vectors in linear manner considering the context of terms.

In traditional NMF (**?**), the high-dimensional matrix $P \in R^{M \times N}$ is approximated by learning two factor matrices $W \in R^{M \times g}$ and $H \in R^{N \times g}$ where g is the number of cluster groups or common topics in the data collection, as follows.

$$P \approx WH^T \tag{1}$$

The matrix factorization process approximates the lower dimensional non-negative factor matrices W and H such that they can represent high dimensional P with the least error. The objective function in Eq 2 iteratively attempts to update the entries in W and H to find the optimum values which possess minimum sum of square error for all the elements in both of those matrices. NMF uses Frobenius norm as its objective function and will find the optimum value of W and H iteratively.

$$\min \frac{1}{2}\|P - WH\| = \sum_{i=1}^{M}\sum_{j=1}^{N}\left(P_{i,j} - (WH)_{i,j}\right)^2 \tag{2}$$

Matrix H contains the information regarding topic membership of each document. Topic membership of each method section in P is obtained considering the maximum coefficient value in H for a method section.

We applied the NMF clustering algorithm to the cleaned dataset, varying the number of clusters from 1 to 50. Clustering solutions were assessed using silhouette score, within-cluster dispersion and between-cluster dispersion.

2.4 Within topic analysis

Within the *PLOS ONE* dataset, statistical methods sections with high topic coherence scores were used to investigate evidence of boilerplate text, compared with other studies assigned to the same topic. Comparisons were carried out by calculating Jaccard similarity scores between studies at the sentence level. We chose the Jaccard score as an intuitive measure, which summarises the similarity between two sentences as the number of words common to each sentence (intersection), divided by the number of words that appeared in either the strongest match or the sentence being compared. Boilerplate text was defined based on word count (plus or minus 3 words) and a Jaccard score of 0.9 or higher. Results were transformed to lower case for the clustering, but examples are given using the original capitalisation.

2.5 Missing statistical methods sections

The statistical methods section for the ANZCTR data was missing for some studies and we examined if there were particular studies where this section was more likely to be missing. We used four independent variables of date, study type (observational or interventional), number of funders and target sample size. 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.

3 Results

3.1 *PLOS ONE*

API searches returned 131,847 results (DOIs) based on chosen search terms. After partial matching, 111,731 (85%) statistical methods sections were identified. In the final sample, 95,518 (85%) DOIs returned an exact match against common section headings: 64,133 for 'statistical analysis', 13,380 for 'statistical analyses' and 13,627 for 'data analysis'. For our random sample of DOIs that did not meet the partial matching criteria, initial search terms appeared in Introduction (number), Results (number) only. xxx appeared in the Methods section without a section heading and xxx had a non-standard section heading. [TODO].

Search results varied by journal volume (Figure 1A). The total number of API search results peaked at volumes 8 (n = 19,045) and 9 (n = 19,045) in years 2013 and 2014. This trend aligned with the total number of papers published in *PLOS ONE* over the same period. 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).

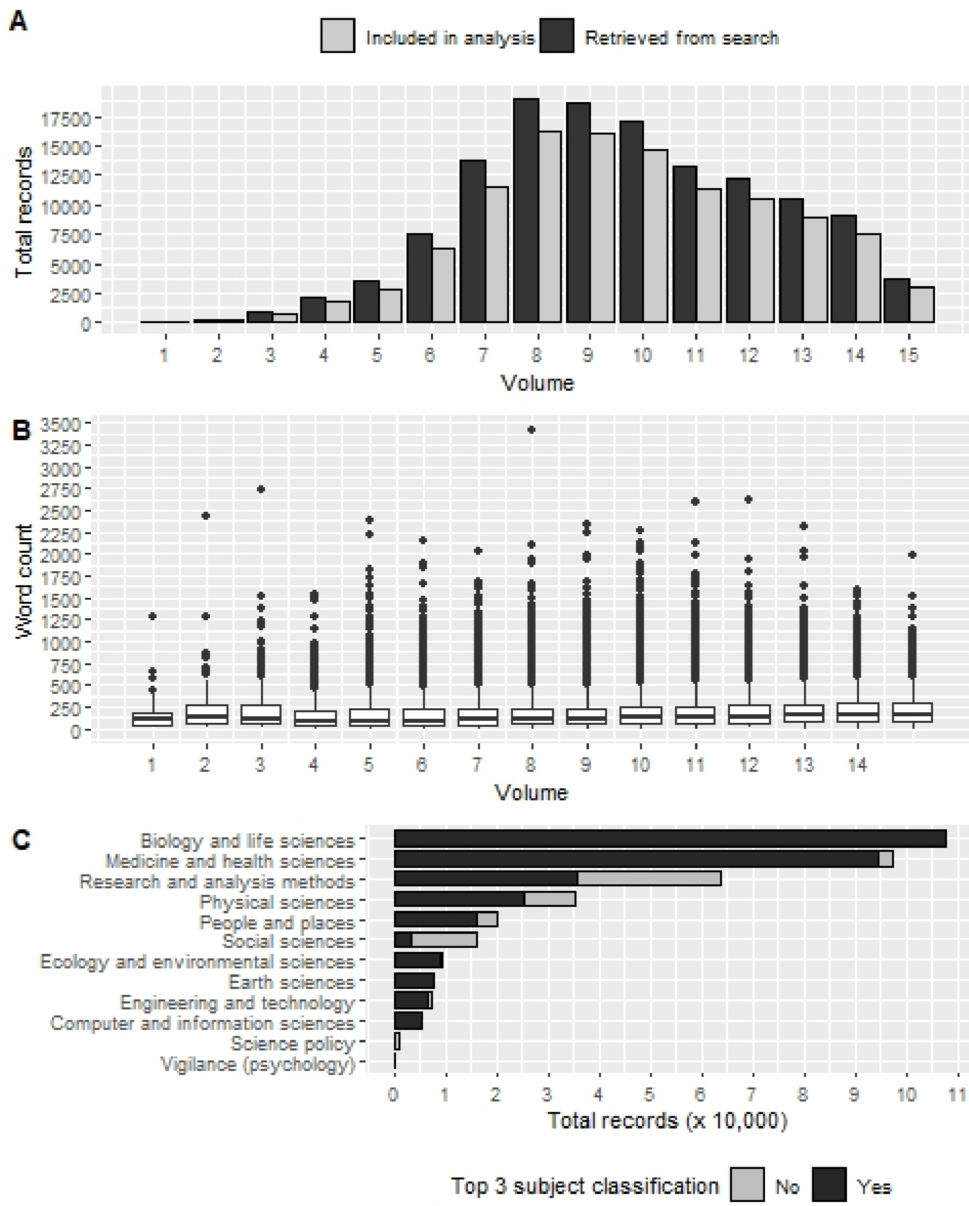

Figure 1: A: Search results by PLOS ONE volume; B: word count per statistical methods section included in analysis (n = 111,73); C: subject classifications assigned to full-text records included in analysis

Statistical methods sections had a median length of 127 words and inter-quartile range of 61 to 254 words (Figure 1B). 7,450 articles (7%) had a statistical methods section of 500 words or more. 19,461 articles (17%) had sections with 50 words or less, equal to the length of this paragraph.

All DOIs included Biology and life sciences (n = 107,584), Earth sciences (n = 7,605) and/or Computer and information sciences (n = 5,190) in their top 3 subject classifications (Figure 1C).

The topic clouds based on ten clusters are in Figure 2. Frequently occurring words reflected the use of statistical software (topics 3 and 5), descriptive statistics (topic 6), group based hypothesis testing (topics 1 and 4) and definitions of statistical significance (topics 1 and 9). Statistical methods sections related to regression (topic 2) and meta-analysis (topic 7) were also identified.

Topics related to statistical software differentiated between Prism GraphPad (topic 3: n = 9974; 8.6%) and SPSS (topic 5: n = 9648; 8.3%) . Manual review of results showed that nine out of the ten top matching DOIs for topic 3 stated the use of Prism GraphPad, but did not specify which statistical methods were used. Top matching sections for topic 5 included information on SPSS version numbers and definitions of statistical significance.

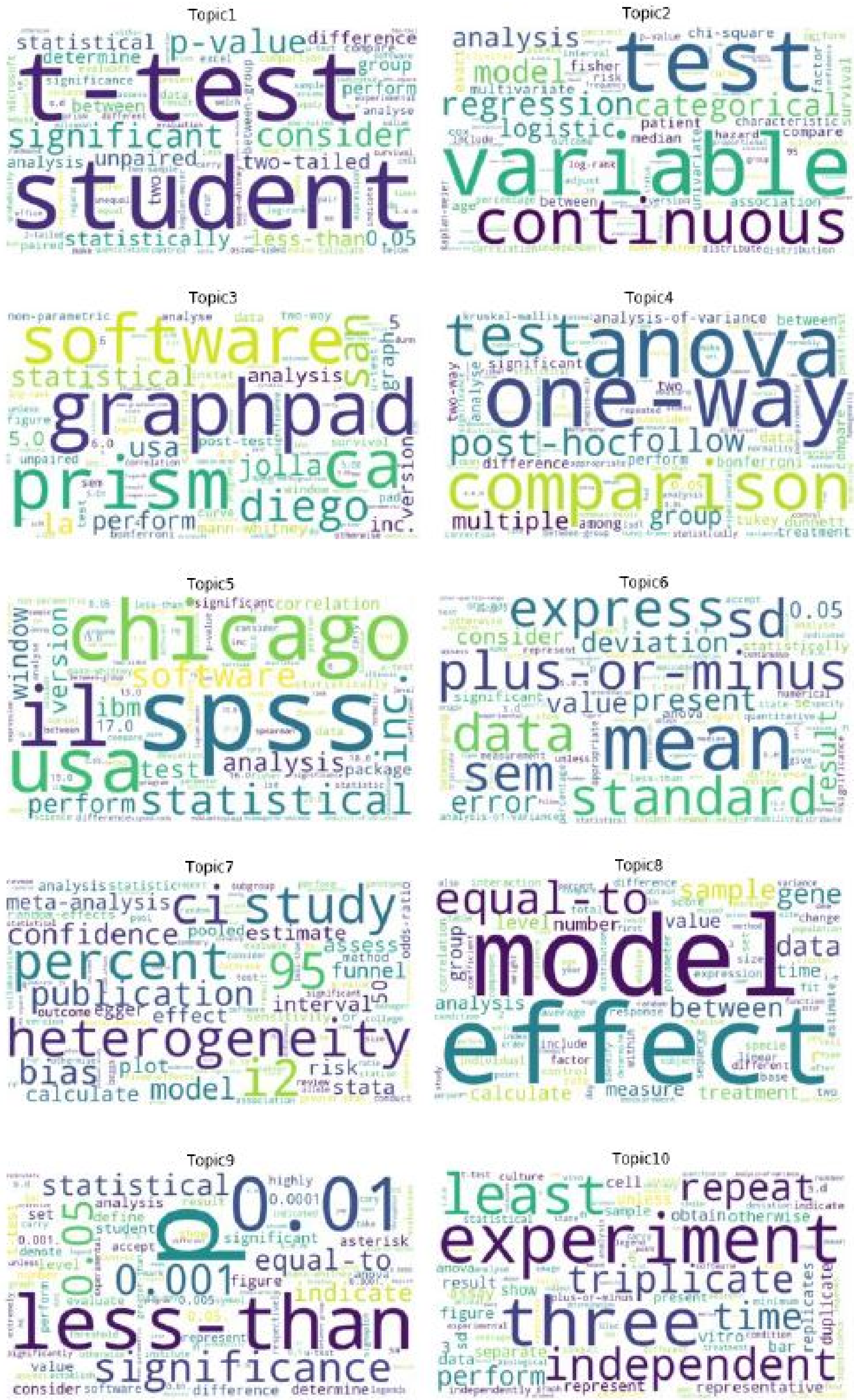

Figure 2: Topic clouds for ten clusters for PLOS ONE

Table 1: Example boilerplate text from PLOS ONE dataset. Sentences marked with an asterisk † correspond to the statistical methods section with the strongest match based on topic coherence score.

Topic	Statistical methods text	Matching DOIs
1	†Students t-test was used for statistical analysis	29
	†A p value of less-than 0.05 was considered statistically significant.	1883
	Statistical analysis was performed using the students t-test	380
	p values of less-than 0.05 were considered significant.	1526
3	†GraphPad Prism (Graphpad Software, San Diego, CA) was used for all analyses.	1941
	All statistical analysis was performed using Graphpad Prism software.	1312
4	†Significant differences were determined using analysis of variance (ANOVA) followed by Tukey post-hoc tests for multiple comparisons.	2216
	Data are expressed as the mean ± SEM	860
	Statistical analysis was performed using one-way analysis of variance (ANOVA) test, followed by Dunnett's multiple comparison post hoc test	4516
	P values less than 0.05 were considered significant	2414
5	†SPSS software version 17.0 (SPSS, Chicago, IL, USA) was used for statistical analysis.	441
	The results are presented as the mean ± SEM.	420
	Data were analyzed by one-way ANOVA and LSD tests using the SPSS 13.0 software (SPSS Inc., Chicago, IL, USA).	19
	The difference was considered statistically significant at P<0.05.	2555
6	†All results are expressed as means ± standard deviation (± SD).	102
	Data are expressed as mean ± Standard error of the mean (SEM)	604
	Statistical analysis of the data was performed by the Student's t-test	444
	A value of P<0.05 was considered statistically significant.	1786
9	†Statistical significance was determined by Students t-tests.	49
	†$^{*}P < 0.05$, $^{**}P =< 0.01$, $^{***}P < 0.001$.	6
	Data are presented as the mean ± SD.	378
	Statistical analysis was performed using the Students unpaired t-test.	752
	Differences were considered statistically significant at $^{*}p < 0.05$, $^{**}p < 0.01$ and $^{***}p < 0.001$.	413

Definitions of statistical significance featured strongly in topic 1 (n = 3784; 3.2%) and topic 9 (n = 6195; 5.3%), combined with text describing hypothesis testing for comparing differences between two groups. Topic 1 reflected applications of Student's t-test assuming a 5% level of statistical significance. Topic 9 referenced similar methods combined with multiple thresholds for declaring statistical significance by asterisk: "$^{*}p < 0.05$, $^{**}p < 0.01$ and $^{***}p < 0.001$", a practice that has been criticised (Wasserstein et al. 2019).

Group-based hypothesis testing was a recurring theme across topics, with text varying based on method(s) used. One-way analysis of variance featured strongly in topic 4 (n = 10212; 8.8%), combined with common methods for performing post-hoc multiple comparisons, for example, Tukey's test and Dunnett's test. Frequently occurring words in topic 6 (n = 4764; 4.1%) reflected mentions of descriptive statistics for summarising continuous variables, for example: "Data are expressed as means ± standard error of the mean (SEM)." Sections assigned to this topic appeared to be expanded versions of topics 1 and 9. Examples of boilerplate text were in the form of descriptive statistics followed by hypothesis testing, for example, Student's t-test, Mann Whitney U or one-way analysis of variance.

Examples of boilerplate text for selected topics are presented in Table 1.

3.2 ANZCTR

We downloaded 28,008 studies. The numbers of excluded studies are shown in Figure 3. Of the 12,700 included studies, 9,523 (75%) had a statistical methods section. The median length of sections was 129 words with an inter-quartile range of 71 to 219 words.

We examined if four study characteristics were associated with a missing statistics section. The odds ratios and 95% credible intervals are in Table 2. 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.

The clustering algorithm found groups that were purely sample size calculations (topic 2, n = 1834; topic 4, n = 909), pilot studies (topic 5, n = 834), safety/tolerability studies (topic 6, n = 524), intervention studies (topic 8, n = 1020) and repeated measures ANOVA

Table 2: Regression results for factors associated with missing statistical methods sections in ANZCTR

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)

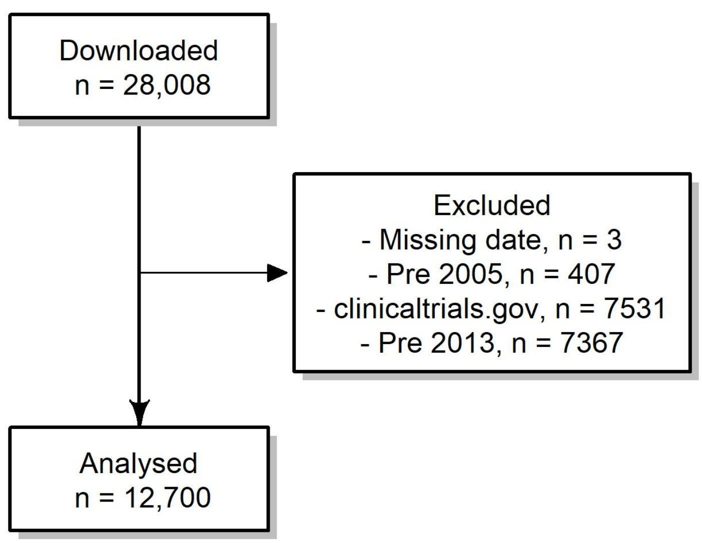

Figure 3: ANZCTR search results.

Figure 4: ANZCTR topic clouds.

(topic 10 = 852). Examples of boilerplate text are provided in Table 3 and Supplementary File 3.

Some methods sections were only one word, including “ANOVA”, “t-test”, “SPSS” and even “SSPS”. There were cases where the exact same method section had been re-used in a different study. For example, in topic 7 (n = 333), 232 sections stated ‘descriptive statistics’ or ‘descriptive statistics used’ with no additional details provided. Other instances with expanded descriptions of methods included topic 3 (n = 1277) outlining descriptive analyses only, and topic 6 (n = 524), for hypothesis testing methods, statistical significance and software.

In other cases, text had been slightly modified to account for changes in primary and secondary outcomes. Examples of these text changes were found in topic 2 (n = 1834) and topic 4 (n = 909); identified instances related to sample size calculations for patient recruitment to different studies (see Supplementary File 3).

Targeted searches for the term ‘statistician’ revealed instances across all topics where no methods were specified. Instead, all analyses were the responsibility of an appointed statistician, for example, “A statistician employed by hospital was used” and “Pilot study at this point will use a statistician professionally to determine sample size calculations as required”.

Table 3: Example boilerplate text from ANZCTR dataset

Topic	**Statistical methods section**
3	Comparisons between categorical variables will be made either using chi square or Fisher exact test. Continuous data will be compared using the Student's t-test or Mann-Whitney U test. Two sided p values of less than 0.05 will be considered statistically significant. The Mann-Whitney U, Student t, 1-way ANOVA, and Kruskal-Wallis tests will be used to compare continuous variables where relevant. The Fisher exact and Pearson's Chi-square test will be used to compare proportions as appropriate.
5	Pilot study No formal sample size calculation was performed
7	Descriptive statistics Descriptive statistics used
9	Linear mixed models will be used to analyse the data.
10	Repeated measures of ANOVA Pre-, during, post- and follow-up variables will be subjected to mixed methods and repeated measures analyses to determine significant changes over (group and) time.

4 Discussion

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 far from the truth, 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.

A frequent theme in the boilerplate statistical methods is the definition of statistical significance, nearly always using a p-value at the 5% level. This widespread use of statistical significance is troubling giving the bright-line thinking it engenders (McShane et al. 2019) and the common misinterpretations of p-values (Goodman 2008).

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 & Goodman 2020).

A related paper has criticised vague statistical methods sections because they deprive readers and reviewers for the opportunity to confirm that the appropriate methods were used (Weissgerber Tracey et al. 2018). These authors checked hundreds of papers using ANOVA and found that 95% did not contain the information needed to determine what type of ANOVA was performed. This lack of information could well be because the authors used a boilerplate statistical methods section that was missing key details.

If authors shared their code then this would provide an alternative route for checking what statistical methods were used. This is not a perfect solution, as we still want authors to accurately report their methods, but it does increase transparency. However, a recent paper found that code sharing was very low in biomedical papers, with just 2% of a sample of over 6,000 papers sharing code (Serghiou et al. 2021).

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 & Saltelli 2018). This is concerning because using the wrong statistical methods can reduce the value of study, or worse, invalidate the entire study. These mistakes are avoidable and are wasting of thousands of hours of researchers' time and the time of patients and volunteers. Poor statistical practice is a key driver of the ongoing reproducibility crisis in science (Ioannidis et al. 2014).

4.1 Limitations

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

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 large 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, D. B., Brown, A. W., George, B. J. & Kaiser, K. A. (2016), 'Reproducibility: A tragedy of errors', *Nature* **530**(7588), 27–29.
URL: *https://doi.org/10.1038/530027a*

Altman, D. G. (1994), 'The scandal of poor medical research', *BMJ* **308**(6924), 283–284.
URL: *https://doi.org/10.1136/bmj.308.6924.283*

Altman, D. G. & Simera, I. (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.
URL: *https://doi.org/10.1177/0141076815625599*

ANZCTR (2019), ANZCTR data field definitions v25, Technical report.
URL: *https://www.anzctr.org.au/docs/ANZCTR%20Data%20field%20explanation.pdf*

Bland, M. (2015), *An Introduction to Medical Statistics*, Oxford medical publications, Oxford University Press.

Brown, A. W., Kaiser, K. A. & Allison, D. B. (2018), ‘Issues with data and analyses: Errors, underlying themes, and potential solutions’, *Proceedings of the National Academy of Sciences* **115**(11), 2563–2570.
URL: *https://www.pnas.org/content/115/11/2563*

Chamberlain, S., Boettiger, C. & Ram, K. (2020), *rplos: Interface to the Search API for 'PLoS' Journals*. R package version 0.9.0.
URL: *https://CRAN.R-project.org/package=rplos*

Diggle, P., Heagerty, P., Liang, K. & Zeger, S. (2013), *Analysis of Longitudinal Data*, Oxford Statistical Science Series, OUP Oxford.

Dobson, A. & Barnett, A. (2018), *An Introduction to Generalized Linear Models*, Chapman & Hall/CRC Texts in Statistical Science, CRC Press.

Ernst, A. F. & Albers, C. J. (2017), ‘Regression assumptions in clinical psychology research practice—a systematic review of common misconceptions’, *PeerJ* **5**, e3323.
URL: *https://doi.org/10.7717/peerj.3323*

Goodman, S. (2008), ‘A dirty dozen: Twelve p-value misconceptions’, *Seminars in Hematology* **45**(3), 135–140.
URL: *https://doi.org/10.1053/j.seminhematol.2008.04.003*

Goodman, S. N., Altman, D. G. & George, S. L. (1998), 'Statistical reviewing policies of medical journals', *Journal of General Internal Medicine* **13**(11), 753–756.
URL: *https://doi.org/10.1046/j.1525-1497.1998.00227.x*

Hardwicke, T. E. & Goodman, S. (2020), 'How often do leading biomedical journalsuse statistical experts to evaluate statistical methods? The results of a survey'.
URL: *osf.io/preprints/metaarxiv/z27u4*

ICJME (2019), 'Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals'.
URL: *http://www.icmje.org/icmje-recommendations.pdf*

Ioannidis, J. P. A., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., Schulz, K. F. & Tibshirani, R. (2014), 'Increasing value and reducing waste in research design, conduct, and analysis', *The Lancet* **383**(9912), 166–175.
URL: *https://doi.org/10.1016/s0140-6736(13)62227-8*

King, K. M., Pullmann, M. D., Lyon, A. R., Dorsey, S. & Lewis, C. C. (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), 547–562.
URL: *https://doi.org/10.1037/abn0000417*

Lang, T. & Altman, D. (2013), Basic statistical reporting for articles published in clinical medical journals: the SAMPL guidelines, *in* P. Smart, H. Maisonneuve & A. Polderman, eds, 'Science Editors' Handbook', European Association of Science Editors.

Leek, J., McShane, B. B., Gelman, A., Colquhoun, D., Nuijten, M. B. & Goodman, S. N. (2017), 'Five ways to fix statistics', *Nature* **551**(7682), 557–559.
URL: *https://doi.org/10.1038/d41586-017-07522-z*

McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. (2019), 'Abandon statistical significance', *The American Statistician* **73**(sup1), 235–245.
URL: *https://doi.org/10.1080/00031305.2018.1527253*

PLOS (2021), Plos one: accelerating the publication of peer-reviewed science, Technical report.
URL: *https://journals.plos.org/plosone/s/criteria-for-publication*

Rinker, T. W. (2018), *textclean: Text Cleaning Tools*, Buffalo, New York. version 0.9.3.
URL: *https://github.com/trinker/textclean*

Serghiou, S., Contopoulos-Ioannidis, D. G., Boyack, K. W., Riedel, N., Wallach, J. D. & Ioannidis, J. P. A. (2021), 'Assessment of transparency indicators across the biomedical literature: How open is open?', *PLOS Biology* **19**(3), 1–26.
URL: *https://doi.org/10.1371/journal.pbio.3001107*

Stark, P. B. & Saltelli, A. (2018), 'Cargo-cult statistics and scientific crisis', *Significance* **15**(4), 40–43.
URL: *https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1740-9713.2018.01174.x*

Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. (2019), 'Moving to a world beyond p ¡ 0.05', *The American Statistician* **73**(sup1), 1–19.
URL: *https://doi.org/10.1080/00031305.2019.1583913*

Weissgerber Tracey, L., Garcia-Valencia, O., Garovic, V. D., Milic, N. M. & Winham, S. J. (2018), 'Why we need to report more than 'Data were analyzed by t-tests or ANOVA'', *eLife* **7**.
URL: *https://gateway.library.qut.edu.au/login?url=https://search.proquest.com/docview/2174217344?*

Wikipedia (2021), Boilerplate text, Technical report.
URL: *https://en.wikipedia.org/wiki/Boilerplate_text*

Zhou, Y. & Skidmore, S. (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.
URL: *https://journals.uair.arizona.edu/index.php/jmmss/article/view/22019*