# 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


April 21, 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 . . .

# 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 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, covering peer-reviewed journal articles and registered clinical trial protocols.

### 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. Articles must be in English. 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 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 et al. 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 in 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 and subject classification(s).

*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 agai

#### 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

We applied the same text cleaning to both data sources. Text cleaning aimed to standardise notation and statistical terminology, whilst minimising changes to article style and formatting. *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 characters corresponding to $\theta$ (<U+03B8>) were replaced with ‘theta’. Similarly, common symbols outside of Unicode blocks including ‘%’ (percent) and ‘<’ (‘less-than’) were converted into plain text using 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 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 (Dobson & Barnett 2018, Diggle et al. (2013),Bland (2015)). The final list is provided as Supplementary Material. 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'.

nst 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…]

## 2.3 Clustering algorithm

*Details to come*

Results were transformed to lower case for the clustering, but examples are given using the original capitalisation.

Results were examined for evidence of boilerplate text in two ways. We first reviewed the top ten results that represented the strongest matches to each cluster. To compare results within a cluster, we calculated cosine similarity score for all pairs of records; higher scores denoted a higher degree of similarity in text between a pair of records.

## 2.4 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 unique records, of which 111,731 (85%) included a statistical methods section based on our search criteria. In the final sample, 95,518 (85%) DOIs returned an exact match against common section headings, including 64,133 for ‘statistical analysis’, 13,380 for ‘statistical analyses’ and 13,627 for ‘data analysis’. Among DOIs that did not meet the partial matching criteria, initial search terms appeared in [TODO].

Search results varied by journal volume (Figure **??**A). The total number of API search results peaked at volumes 8 (n = 19,045) and 9 (n = 19,045), corresponding to 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).

The median length of statistical methods sections was 127 words (IQR: 61 to 254 words) (Figure **??**B). 7,450 articles (7%) had a statistical methods section of 500 words or more. 19,461 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 = 107,584), Earth sciences (n = 7,605) and/or Computer and information sciences (n = 5,190) in their top 3 subject classifications (Figure **??**C).

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.

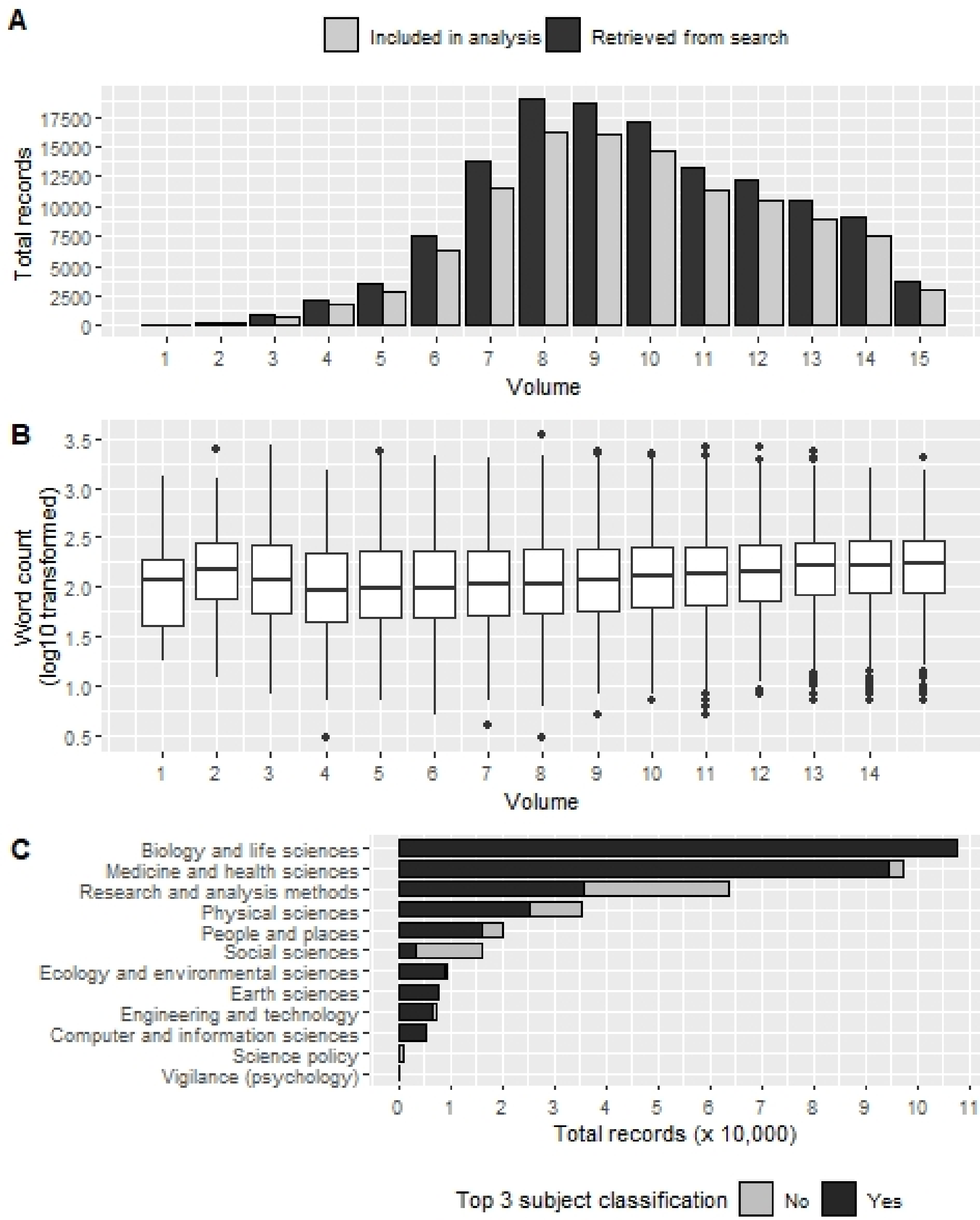


Figure 1: PLOS ONE result summary.

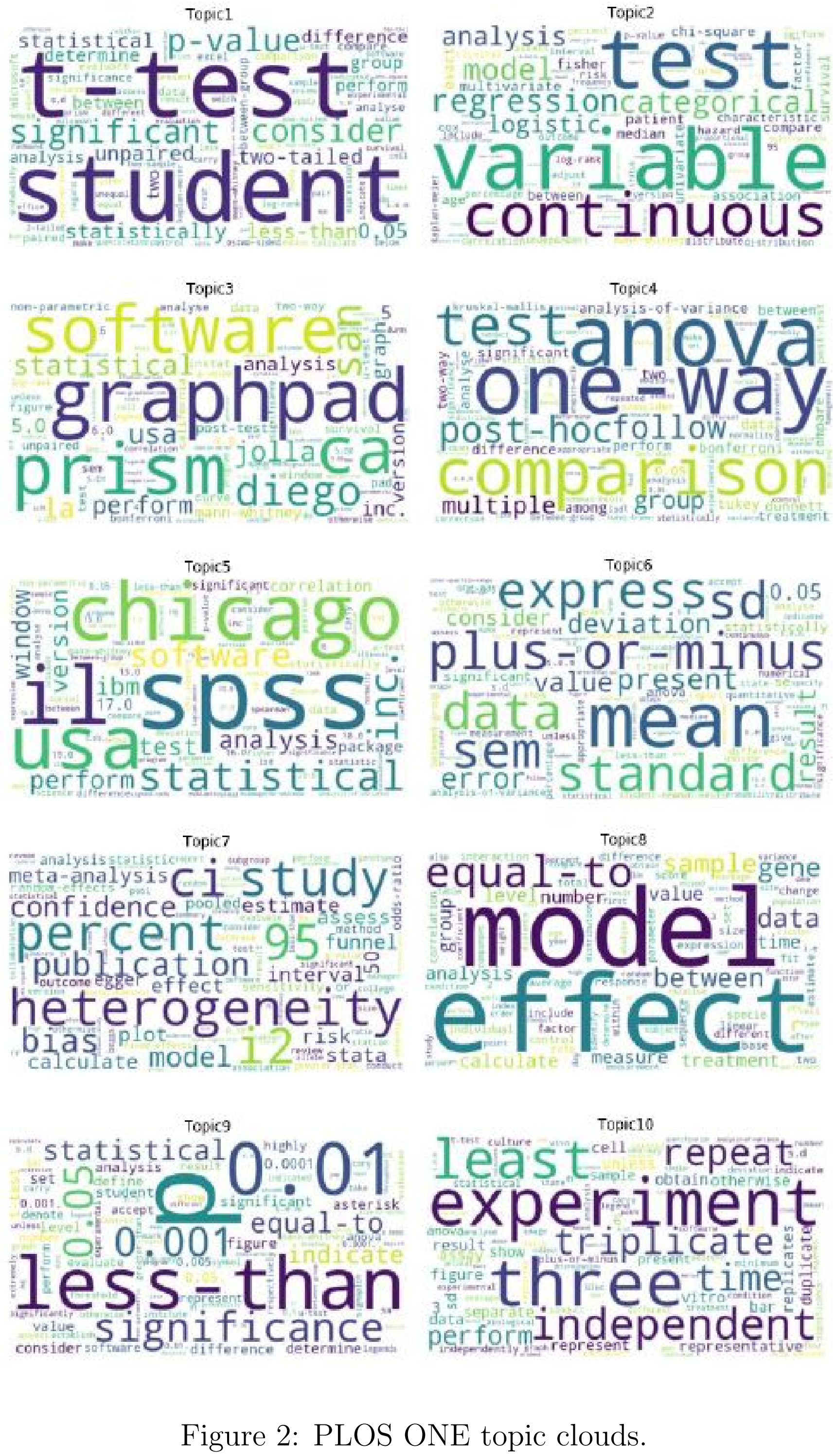


Figure 2: PLOS ONE topic clouds.

| topic_id | DOIs | Median word count | Q1 | Q3 |
|---|---|---|---|---|
| Topic 1 | 3784 | 36 | 27 | 52 |
| Topic 10 | 4697 | 56 | 43 | 79 |
| Topic 2 | 19658 | 181 | 120 | 266 |
| Topic 3 | 9976 | 61 | 44 | 88 |
| Topic 4 | 10216 | 69 | 49 | 105 |
| Topic 5 | 9652 | 74 | 52 | 112 |
| Topic 6 | 4768 | 46 | 36 | 63 |
| Topic 7 | 5307 | 199 | 129 | 290 |
| Topic 8 | 42591 | 287 | 177 | 450 |
| Topic 9 | 6211 | 56 | 39 | 84 |

The topic clouds based on ten clusters are in Figure˜**??**. 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). There are also statistical methods sections associated with regression (Topic˜2) and meta-analysis (Topic˜7).

Topics related to the use of statistical software differentiated between Prism GraphPad (Topic˜3: n = 9,879; 8.8%) and SPSS (Topic 5: n = 9,574; 8.6%) (Box 1). A manual review of the top matching sections in these topics showed strong evidence of boilerplate text. Nine out ten top matches for Topic 3 stated the use of Prism GraphPad but did not specify which statistical methods were used; six out of ten top matches returned the same cluster score indicating near identical text. Top matching sections for Topic 5 included information on SPSS version numbers and definitions of statistical significance.

### 3.2 ANZCTR

## 4 Supplementary

### 4.1 PLOS


Figure 3: ANZCTR topic clouds.

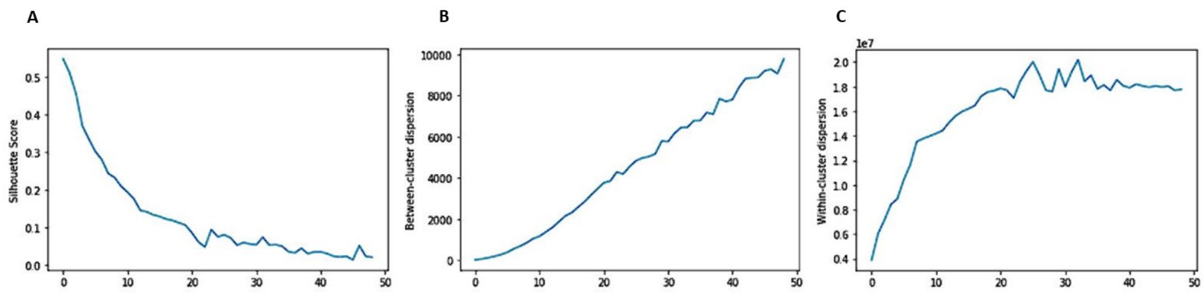


Figure 4: PLOS ONE clustering metrics.

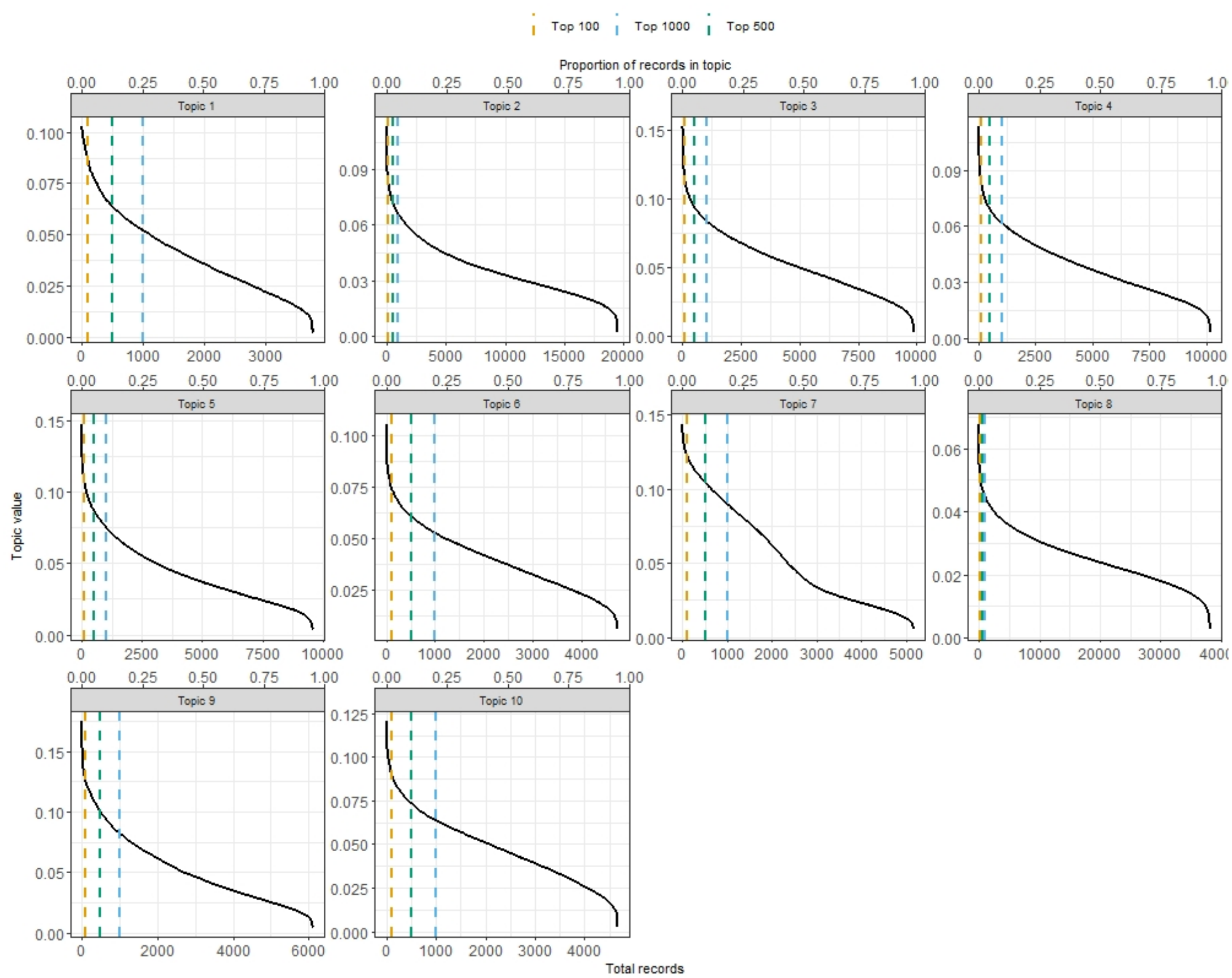


Figure 5: PLOS ONE topic values.

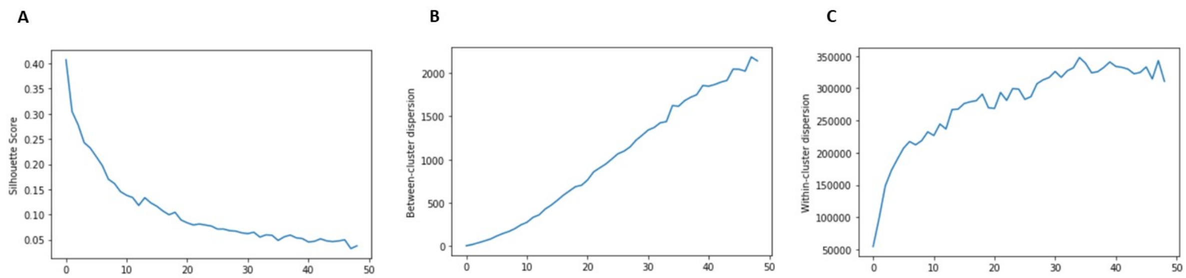


Figure 6: ANZCTR clustering metrics.

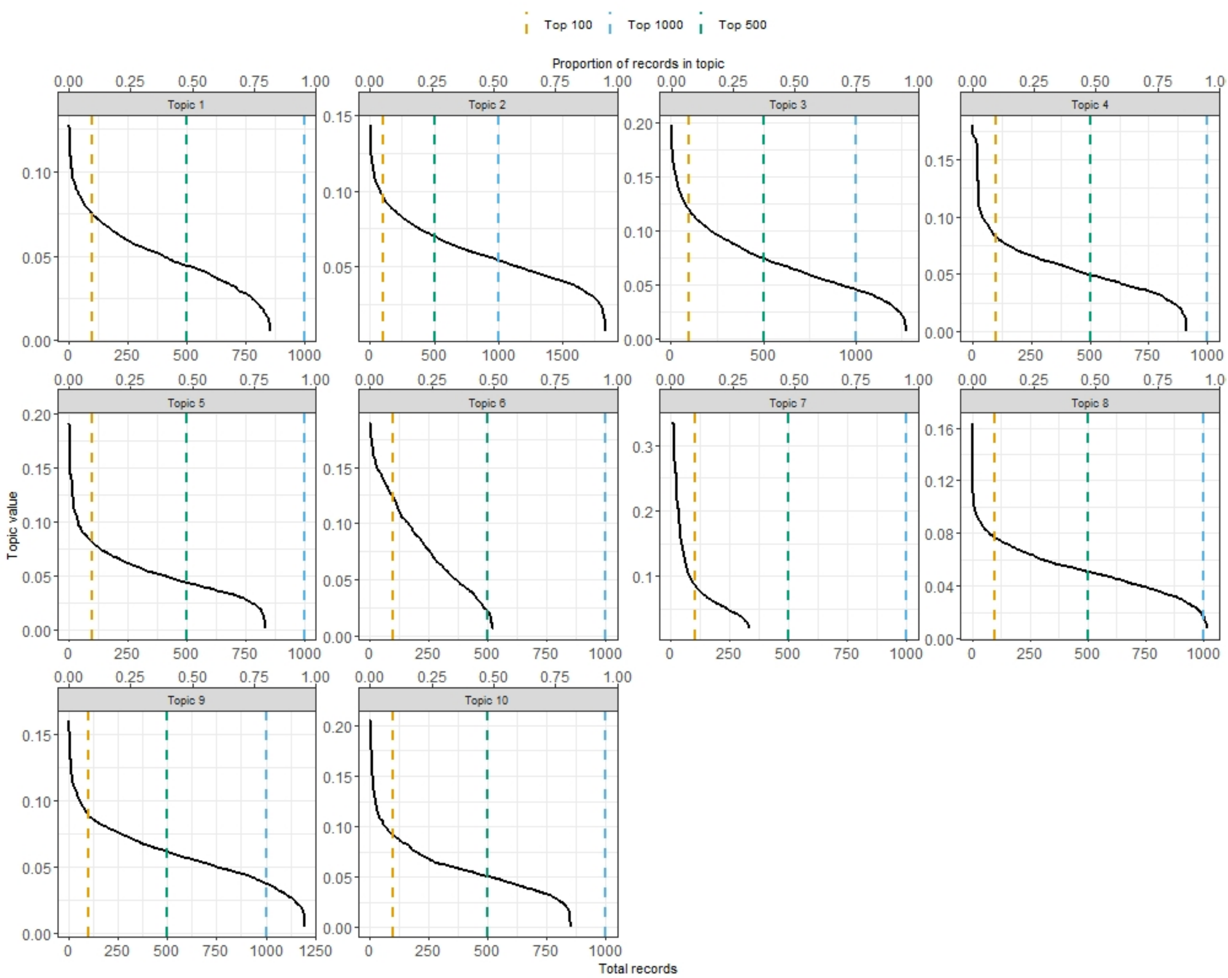


Figure 7: ANZCTR topic values.

### 4.2 ANZCTR

## 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. 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*

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*

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*

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*

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*