ROYAL SOCIETY OPEN SCIENCE

Comparing the prevalence of statistical reporting inconsistencies in COVID-19 preprints and matched controls: A Registered Report

Journal:	*Royal Society Open Science*
Manuscript ID	RSOS-202326.R2
Article Type:	Registered Report - Stage 1
Date Submitted by the Author:	23-Feb-2021
Complete List of Authors:	van Aert, Robbie C. M.; Tilburg University, Nuijten, Michèle B.; Tilburg University, Department of Methodology and Statistics Olsson-Collentine, Anton; Tilburg University, Methodology and Statistics Stoevenbelt, Andrea H.; Tilburg University, Methodology and Statistics van den Akker, Olmo; Tilburg University, Methodology and Statistics Wicherts, Jelte ; Tilburg University,
Subject:	Statistics < MATHEMATICS
Keywords:	statistical inconsistencies, COVID-19, preprint, registered report
Subject Category:	Mathematics

Comparing the prevalence of statistical reporting inconsistencies in COVID-19 preprints and matched controls: A Registered Report

Robbie C. M. van Aert,* Michèle B. Nuijten,* Anton Olsson-Collentine, Andrea H. Stoevenbelt, Olmo van den Akker, Jelte M. Wicherts

* Shared first authorship

Abstract

The COVID-19 outbreak has led to an exponential increase of publications and preprints about the virus, its causes, consequences, and possible cures. COVID-19 research is performed under high time pressure and is subject to financial and societal interests. Doing research under such pressure may affect the scrutiny with which researchers perform and write up their studies. In line with this notion, preliminary evidence shows signs of suboptimal methodology and reporting in COVID-19 research. Such "high-speed" science may also lead to errors in the reporting of statistical results. Incorrectly reported statistics can seriously affect the validity of conclusions. In this project, we plan to manually extract a range of statistics from 1,200 COVID-19 and 1,200 non-COVID-19 preprints, check their internal consistency, and compare the prevalence of statistical reporting inconsistencies. We hypothesize a difference in statistical reporting inconsistencies between COVID-19 and non-COVID-19 preprints.

The COVID-19 pandemic has led to an exponential increase of publications and preprints (i.e., scientific manuscripts published in open registries that have yet to undergo peer review) concerning COVID-19 (Brainard 2020; Callaway, 2020). Not only the volume of output is almost unprecedented, COVID-19 related articles are also published much faster than their non-COVID-19 counterparts (Horbach, 2020). It is encouraging to see the speed with which the scientific community is responding to this pandemic, but this "high-speed" science may not be without risks. A substantial number of scientists have voiced their concern that the pressure to disseminate findings quickly may decrease scrutiny in performing, reporting, and reviewing COVID-19 studies (Horbach, 2020; Marcus & Oransky, 2020; Dinis-Oliveira, 2020; Peyrin-Biroulet, 2020; Bramstedt, 2020; Dobler, 2020).

Several empirical studies underline the possible risks of publishing science under pressure. For example, past bibliometric analyses showed low international collaboration in research on natural disasters (incl. the severe acute respiratory syndrome coronavirus [SARS-CoV] outbreak in 2003; Chiu et al., 2004; Sweileh, 2019; Lee & Lin, 2020). Furthermore, analyses of the current COVID-19 literature showed poor methodology and reporting in both preprints and peer-reviewed publications (e.g., small samples, lack of control groups, no mention of a sampling frame or study limitations; Alexander et al., 2020; Glasziou et al., 2020; Ramanan et al., 2020; Weissgerber et al., 2020). In a more direct comparison, a recent preprint reported lower methodological rigor in 539 published COVID-19 papers compared with papers published in the previous year in the same journal as the control group (Jung et al., 2020). Specifically, Jung et al. (2020) measured methodological rigor using standard quality checklists such as the Cochrane Risk of Bias Tool (Higgins et al., 2011) and the Newcastle-Ottawa Scale (Wells et al., 2021). They showed that only 41% of the COVID-19 articles could be considered of high methodological quality, as compared to over 73% of the articles in the control group.

To our knowledge, one important aspect of research quality of COVID-19 studies has not been looked at yet: the quality of statistical reporting. Statistics underlie many conclusions presented in COVID-19 studies, so it is of paramount importance that these statistics are correctly reported. Inconsistent statistics affect the robustness of a conclusion: if a reported statistic is not in line with other information in the paper, the trust in the claim will be lowered. Imagine, for example, that researchers study the impact of a treatment on the mortality of patients infected with the virus. The researchers report an odds ratio, but the odds ratio is not in line with the number of patients in the experimental and control group that die. Such an inconsistency will lower the trust in the finding and corresponding substantive conclusions and policy.

Previous research has shown a high prevalence of statistical reporting inconsistencies in publications from different fields (Georgescu & Wren, 2018; Brown & Heathers, 2017; Nuijten et al., 2016; Petrocelli et al., 2013; Bakker & Wicherts, 2011). Based on the observed suboptimal methodological and reporting quality of COVID-19 studies, and given the increased time pressure and strong financial and societal interests under which this research is conducted, it is imaginable that the prevalence of statistical reporting inconsistencies in COVID-19 papers is even higher than in non-COVID-19 papers. However, the contrary could also be true: the severity of the pandemic, the perceived importance of COVID-19 research, and the expected attention this research may cause researchers to take more care in reporting the statistical results than they would normally do. To study this, we will compare the prevalence of statistical reporting inconsistencies in COVID-19 preprints and matched controls. Specifically, we will test the following hypothesis:

Hypothesis: The prevalence of statistical reporting inconsistencies differs between COVID-19 and matched non-COVID-19 preprints.

We specifically focus on preprints for two main reasons. First, preprints have played a central role in early dissemination of scientific insights in the current pandemic (Callaway, 2020). Second, preprints are easy to access because they are not behind paywalls and the majority of preprints are published at dedicated preprint servers, which means that they can easily be located and accessed.

Methods

We will study the prevalence of statistical reporting inconsistencies in COVID-19 as compared to non-COVID-19 preprints. This makes our study a natural experiment, because we are comparing two existing groups that occurred naturally and cannot be controlled. A summary of the design can be found in Table 1. All data and code for this project will be available on the Open Science Framework (OSF) on https://osf.io/tdfgq/.

Population of preprints under study

The preprint servers medRxiv and bioRxiv collate all submitted preprints about COVID-19 research (see https://connect.biorxiv.org/relate/content/181). The population of cases that we will study are the preprints that are published on medRxiv and bioRxiv and classified as COVID-19 research by these servers. Our sampling frame consists of all COVID-19 preprints that were available from the preprint servers between January 19, 2020 and January 31, 2021. January 19, 2020 was selected as the starting date, because this is the date of the oldest COVID-19 preprint on the servers (accessed on February 2, 2021). The sampling frame also contains information about characteristics of the preprints, such as the number of authors, the subject category, the server a preprint was published on, and the date a preprint was published. These four characteristics will be used for drawing a stratified random sample of 1,200 COVID-19 preprints (see the sample size section) to ensure that the characteristics of the population of preprints are represented as closely as possible in the sample.

The strata used in this sampling procedure will be: the preprint server, (medRxiv or bioRxiv), subject category as identified by the preprint servers, month of the year a preprint was published, and a categorical variable indicating the number of authors of a preprint (categories 1, 2, 3-10, 11-∞). We will include strata for preprint server and subject category as the prevalence of statistical reporting inconsistencies may differ between research areas. Month of the year will be included to take potential differences into account between preprints published at the start of the pandemic or more than one year later. The categories for the number of preprint authors are selected to reflect possible differences in the sense of responsibility authors may feel to double-check the reported statistical results. It is imaginable that there is a difference between single-authored and multi-authored preprints, because co-authors may check the statistical results in multi-authored preprints. It is also imaginable that there is a diffusion of responsibility as the number of authors increases, which is reflected by the different categories for 2, 3-10, and 11-∞ authors.

We also have a sampling frame of all non-COVID-19 preprints published between January 19, 2020 and January 31, 2021. The same strata as above will be used to match each sampled COVID-19 preprint to a comparable non-COVID-19 preprint. That is, for each COVID-19 preprint we will select a non-COVID-19 preprint that is published on the same preprint server, in the same subject category, in the same month of the year, and with the same category of number of authors. The most recent version of a COVID-19 preprint was always downloaded, and we will also include the version number of the COVID-19 preprint for selecting a matching non-COVID-19 preprint. If multiple non-COVID-19 preprints end up in the selection, we will randomly sample one of these preprints. We will only search for a matching preprint if the corresponding COVID-19 preprint contains statistical results of which we could check the consistency. In case a COVID-19 preprint contains statistical results but the matching preprint does not, we will continue randomly sampling a matching preprint until we have sampled one that contains statistical results.

Table 1. Design table.

Question	Hypothesis	Sampling plan (e.g., power analysis)	Analysis Plan	Interpretation given different outcomes
Does the prevalence of statistical reporting inconsistencies differ between COVID-19 and non-COVID-19 preprints?	The prevalence of statistical reporting inconsistencies differs between COVID-19 and non-COVID-19 preprints	Our resources allow us to study 1,200 COVID-19 preprints and 1,200 non-COVID-19 preprints. This gives us 80% power to detect an effect if the odds ratio of a statistical inconsistency in a COVID-19 preprint versus a non-COVID-19 preprint is about 1.38. See the power analysis section for more details.	We will analyze the data using two logistic multilevel models. The first model only contains the predictor of interest (COVID-19 or not) and the second model also includes relevant control variables. The hypothesis will be tested in both models using a two-tailed test with $\alpha = .05$ and a Bayes factor. See the analysis section for more details.	We will report the odds ratio (and corresponding confidence interval) that describes how much larger (or smaller) the odds are for reporting a statistical inconsistency in a COVID-19 preprint compared to a non-COVID-19 preprint. In case the null-hypothesis cannot be rejected in the frequentist hypothesis test, we will conclude that there is no evidence for a difference in the prevalence of statistical reporting inconsistencies between COVID-19 and non-COVID-19 preprints. The Bayes factor will be interpreted with a statement along the lines of: “the model where the prevalence in statistical reporting inconsistencies differs for COVID-19 and non-COVID-19 preprints is X times more likely than the model where the prevalence is the same for both COVID-19 and non-COVID-19 preprints”

In case there are no matching non-COVID-19 preprints for a certain COVID-19 preprint, we will relax the month in which a preprint was published by also searching for matching preprints in adjacent months. If there are also no non-COVID-19 preprints published with these characteristics in adjacent months, we will again search for preprints published in the same month as the COVID-19 preprint but now relaxing the characteristic reflecting the number of authors by assessing whether a non-COVID-19 preprint is present in adjacent categories. Third, if this did not result in a matching non-COVID-19 preprint, we will search for preprints in the same category reflecting the number of authors in the COVID-19 preprint but published on the other preprint server. Finally, in the rare cases where this will not result in a match, we will look for a non-COVID-19 preprint with the same subject category and version number, but relax all other characteristics. To verify that the matching procedure works as planned, we will manually check whether a non-COVID-19 preprint is indeed a matched case for the first 50 sampled preprints.

We programmed the stratified sampling procedure and all analyses in the statistical software R (version 4.0.3; R Core Team, 2020). R code for the stratified sampling procedure and matching procedure is available at https://osf.io/bmkew/ and https://osf.io/6rhu9/.

Data extraction

The dependent variable in our study is the internal consistency of a reported statistic. A statistical reporting inconsistency arises when numbers belonging to a set do not match. For example, when a paper states "7% of the patients died in the hospital (5/100)", it is clear that these numbers are not internally consistent: 5/100 is 5%, not 7%. In our study, we will manually extract[1] and subsequently assess the internal consistency of the following types of statistics:

- Reported percentages should match the accompanying fraction (e.g., 5% vs. 5/100).
- Reported test sensitivity/specificity/accuracy/positive predictive value/negative predictive value should match reported true positive/true negative/false positive/false negative rates
- Reported total sample size should match reported subgroup sizes
- Reported marginal values in frequency tables should match cell values
- Reported *p*-values should match reported test statistics and degrees of freedom
- Reported odds ratios/risk ratios/risk differences should match values reported in the associated frequency table

We will extract reported statistics regardless of whether these come from a primary or secondary analysis. However, we will not extract reported statistics if the preprint contains insufficient information to assess the internal consistency. We will also not extract other types of statistics than the ones mentioned above (e.g., effect sizes such as explained variance in an ANOVA or regression analysis and results of Bayesian analyses), because these types of statistics are usually not reported with enough detail to allow a check for internal consistency. If more than two tables are reported in a preprint, we will randomly select two to potentially extract statistics from. We will not extract statistics reported in appendices or supplemental materials. See our coding protocol (https://osf.io/y8x4a/) for details.

To check and improve interrater reliability of the coding procedure, the two research assistants responsible for statistics extraction will both code the same 50 preprints. We will assess interrater reliability by verifying that both coders have extracted the same statistics from the preprints. Specifically, we will do two things. First, we will calculate the correlation between the number of statistics extracted by each coder. Second, we will

[1] Note the tool statcheck (Epskamp & Nuijten, 2014) that automatically extracts null hypothesis significance tests and checks the consistency of reported *p*-values cannot be used here, because it can only detect statistics reported in APA style, which is not used in the majority of preprints submitted to bioRxiv and medRxiv.

calculate the median difference between the number of extracted statistics of each coder per preprint. Moreover, a meeting will be organized with the two research assistants and one of the principal investigators to evaluate the data extraction procedure for these 50 preprints. Differences in extracted statistics by the two research assistants will be discussed in this meeting in order to improve interrater reliability.

The extracted data will be filled out in a spreadsheet (https://osf.io/u6f3s/), and we will run an R script (https://osf.io/hnau2/ and https://osf.io/k4w7q/) to automatically recompute the reported statistic from the other reported information. If the reported and recomputed statistics do not match, the result is marked as inconsistent. We will take rounding into account[2] and only compare the statistics after rounding using the same number of decimals as the reported statistic. If the R script flags a result as an inconsistency, we will manually verify this in the preprint to decrease the probability that we wrongly classify statistics as inconsistencies.

In this operationalization of statistical reporting inconsistencies, we will not take into account the size of the discrepancy, because it is often not possible to determine which of the reported results in the set is/are incorrect. To illustrate, say that a paper reports the following statistic: "$t(28) = 2.2$, $p = .063$". The recalculated p-value based on test statistic and degrees of freedom is .036. This could be considered a relatively large discrepancy: the recalculated p-value is almost twice as small as the reported p-value, and it falls on the other side of the conventional significance level of .05. However, without the raw data we have no way of knowing which of the three reported results is incorrect. It could also be the case that the p-value is correctly reported, but there is an error in the test statistic and the recalculated test statistic is 1.9 instead of 2.2. This discrepancy is much smaller (percentage-wise) than the discrepancy in the recalculated p-values, even though we looked at the same set of inconsistent numbers. Therefore, we decided to only record whether a statistic is consistent or inconsistent, and not try to determine the size of the discrepancy.

Analysis

We will present descriptive results of the extracted statistics separately for COVID-19 and non-COVID-19 preprints including: the types of statistics, the number of statistics per preprint, and the number of (in)consistent statistics (these will also be reported per subject category). We will also report the proportion of COVID-19 preprints that contained statistics that we could extract.

We will test our hypothesis using a logistic multilevel model, because the dependent variable in our study is dichotomous (whether a statistical result is inconsistent or not). This model also takes into account any dependencies between inconsistencies in results within a single preprint. We indicate a statistical result with i and a preprint with j such that the statistical model to test our hypothesis is:

$$logit(\pi_{ij}) = \gamma_{00} + \gamma_{01} covid + \mu_{0j}$$

where π_{ij} is the probability of an inconsistency in statistical result i of preprint j, γ_{00} denotes the intercept, γ_{01} is the parameter of interest testing whether COVID-19 preprints contain more or fewer statistical inconsistencies than non-COVID-19 preprints (i.e., $covid$ indicates whether the preprint was about COVID-19 (1) or not (0)), and μ_{0j} denotes the random effect that follows a normal distribution with mean 0 and variance τ^2. Our primary interest is estimating γ_{01} and testing whether this parameter is different from zero, which would indicate that results are either more or less likely (instead of equally likely) to be inconsistent if they are

[2] Different software packages sometimes maintain different rounding rules. Specifically, depending on which software is used for rounding (or if done manually), a value ending in 5 can be either rounded up or down. To take this into account, we will round recalculated values ending in 5 in both directions and if either matches the reported value consider it correct.

reported in a COVID-19 preprint than in a non-COVID-19 preprint. Parameter estimates together with confidence intervals will be reported, and we will test our hypothesis with significance level α = .05. We selected α = 0.05, because we believe a 0.05 probability of making a Type-I error is acceptable in this study. Furthermore, a significance level of α = 0.05 in combination with our sample size enables us to detect an odds ratio of approximately 1.38 with 80% power (see Figure 1), which we consider practically relevant.

Next to conducting a frequentist hypothesis test, we also employ Bayesian hypothesis testing. We will compute approximated adjusted fractional Bayes factors (Gu, Mulder, & Hoijtink, 2017) to compare two models with each other where $\gamma_{01} = 0$ and $\gamma_{01} \neq 0$ using the default implementation in the R package BFpack (Mulder, Williams, Gu, et al., in press). Comparing these two models is the Bayesian equivalent of the two-tailed frequentist hypothesis test that we propose. The approximated adjusted fractional Bayes factor uses a minimal fraction of the available data to train a non-informative normally distributed prior and approximate the marginal likelihood of the tested hypotheses.

We will also run another logistic multilevel model where we extend the model above by including control variables to study whether the estimated effect is affected by including other variables in the model. All control variables will be measured at the preprint level. The control variables that will be included are: the number of authors of a preprint (continuous variable), the number of days a preprint was published after the first COVID-19 preprint was published (January 19, 2020), and the total number of extracted statistics in a preprint. The number of authors of a preprint and number of days a preprint was published after January 19, 2020 were included for the same reasons as using these variables for creating strata in the stratified random sampling procedure. The total number of extracted statistics in a preprint was included to take into account that a statistical inconsistency is more likely to occur if many statistics are reported in a preprint. We will again test whether there is a difference in how likely a statistical reporting inconsistency is in a COVID-19 preprint compared to a non-COVID-19 preprint using the outlined frequentist hypothesis test above (with α = .05) as well as by computing the proposed Bayes factor. R code of our planned analyses is available at https://osf.io/a8ezq/.

Power analysis

For this project, we received €10,000 seed funding from the Herbert Simon Research Institute of Tilburg University. This covers the costs of hiring research assistants for 400 hours who can assist us with collecting data. Based on pilot studies, we estimate that it will take approximately 10 minutes on average to manually extract all the statistical results from a preprint. This means that with our budget, we can screen 2,400 preprints: 1,200 COVID-19 preprints and 1,200 non-COVID-19 preprints. We will collect as many observations as our resources allow. We ran a statistical power analysis to determine how large the population effect must be to detect it with our design and planned analysis. The R code for this analysis can be found on OSF (https://osf.io/6py9x/).

In order to calculate power, we needed 1) a value for γ_{00} (the base rate prevalence of inconsistencies in non-COVID-19 preprints), 2) a value for γ_{01} (the estimated "effect" of a COVID-19 preprint; in other words, how much does the probability that a statistic is inconsistent differs if it is published in a COVID-19 preprint), and 3) a value for τ^2 (the variance of the random effect μ_{0j}). For γ_{00} we assumed a log odds value of -2.761187, based on the base rate prevalence of statistical reporting inconsistencies in psychology articles published in Frontiers in Psychology and PLOS One (Nuijten et al., 2017; Study 2). For γ_{01}, we assumed a range of ten evenly spaced (log odds) values, ranging from 0 (no effect) to 0.758 (equivalent to twice as many inconsistencies in a COVID-19 preprint compared to a non-COVID-19 preprint). Note that we calculate power to detect a positive effect (i.e., a higher probability of inconsistencies in COVID-19 preprints), but the same power estimates would hold for negative effects of the same size. For τ^2 we assumed two values: 2.35 and 1.85, which correspond to the observed τ^2 values in Study 2 and 3 in Nuijten et al. (2017), respectively.

Figure 1 shows statistical power based on 1,000 repetitions for a scenario in which we can screen 1,200 COVID-19 and 1,200 non-COVID-19 preprints, assuming that approximately 45% of the preprints contain statistics (estimated based on pilot rounds of coding). The *x*-axis shows a range of possible population effects: from no effect to a large effect where the error-prevalence is twice as high in COVID-19 preprints. The *x*-axis labels are divided in three rows. Row 1 shows the value of γ_{01}, and rows 2 and 3 show the predicted prevalence of statistical inconsistencies in a given COVID-19 preprint in log odds and in probabilities, respectively. The two lines in the figure represent the results for different values of τ^2 (variance in intercepts). The results show that we would obtain a statistical power of 80% if the true effect (γ_{01}) lies between 0.30 and 0.34. In other words, if the odds ratio of a statistical result being inconsistent in a COVID-19 compared to a non-COVID-19 preprint is approximately 1.38 (i.e., γ_{01} = 0.32; or, corresponding to γ_{01} = -0.32, an odds ratio of 0.73), we should be able to detect it with our design with 80% power. This corresponds to either an increase in the probability that a statistic is inconsistent from .063 to .087 or a decrease from .063 to .046 in non-COVID-19 versus COVID-19 preprints, respectively.

Limitations

Our proposed design comes with certain limitations. First, we specifically focus on the internal consistency of reported statistics. This means that we can only check results for which all relevant information is reported. However, it is imaginable that when the statistical analyses are done less carefully, the reporting may also be done in a similar fashion and not all information might be there to allow a check for consistency. Relatedly, with the proposed procedure we do not spot mistakes or problems in the way the raw data were collected, processed, or analyzed. However, we believe that spotting statistical reporting inconsistencies can shed light on a basic quality requirement of a study, and is therefore interesting on its own.

Second, our population under study includes preprints published on bioRxiv and medRxiv. This comes with two limitations with respect to the generalizability of our results. First, we do not know to what extent our findings will generalize to published papers. An important difference between preprints and published papers is that the latter are peer reviewed. Theoretically, one could expect that this implies that the overall inconsistency rates in published papers will be lower than in unreviewed preprints (see, e.g., Carneiro et al., 2020, who found a small improvement in reporting quality in peer-reviewed papers compared to the respective preprints). To some extent, we might be able to compare the inconsistency rates in the preprints currently under study and the inconsistency rates in published papers as observed in previous research (Georgescu and Wren, 2018), but this comparison will per definition be suboptimal since the approach in previous research and ours are not fully equivalent. Second, we only include preprints from bioRxiv and medRxiv, because these servers provided a curated list of COVID-19 research. It is possible that preprints published on other preprints servers differ from the included preprints in systematic ways. However, since bioRxiv and medRxiv (the latter, specifically) have played the biggest role in distributing COVID-19 preprints (Callaway, 2020), we think they are the most relevant servers to study.

Third, one could argue that the presence of statistical inconsistencies in preprints is not particularly problematic, since preprints are not yet the final publication of the research. However, we would argue against this for two main reasons. First, in the COVID-19 pandemic, preprints have played an important role in the speedy distribution of new knowledge (Callaway, 2020; Weissgerber et al., 2020), so it is important that the information presented in these preprints is correct. Second, it may not be the case that the peer review process is sufficiently able to correct statistical reporting errors present in the preprint, given the high prevalence of statistical reporting inconsistencies in the published literature (Nuijten et al., 2016; Georgescu and Wren, 2018).

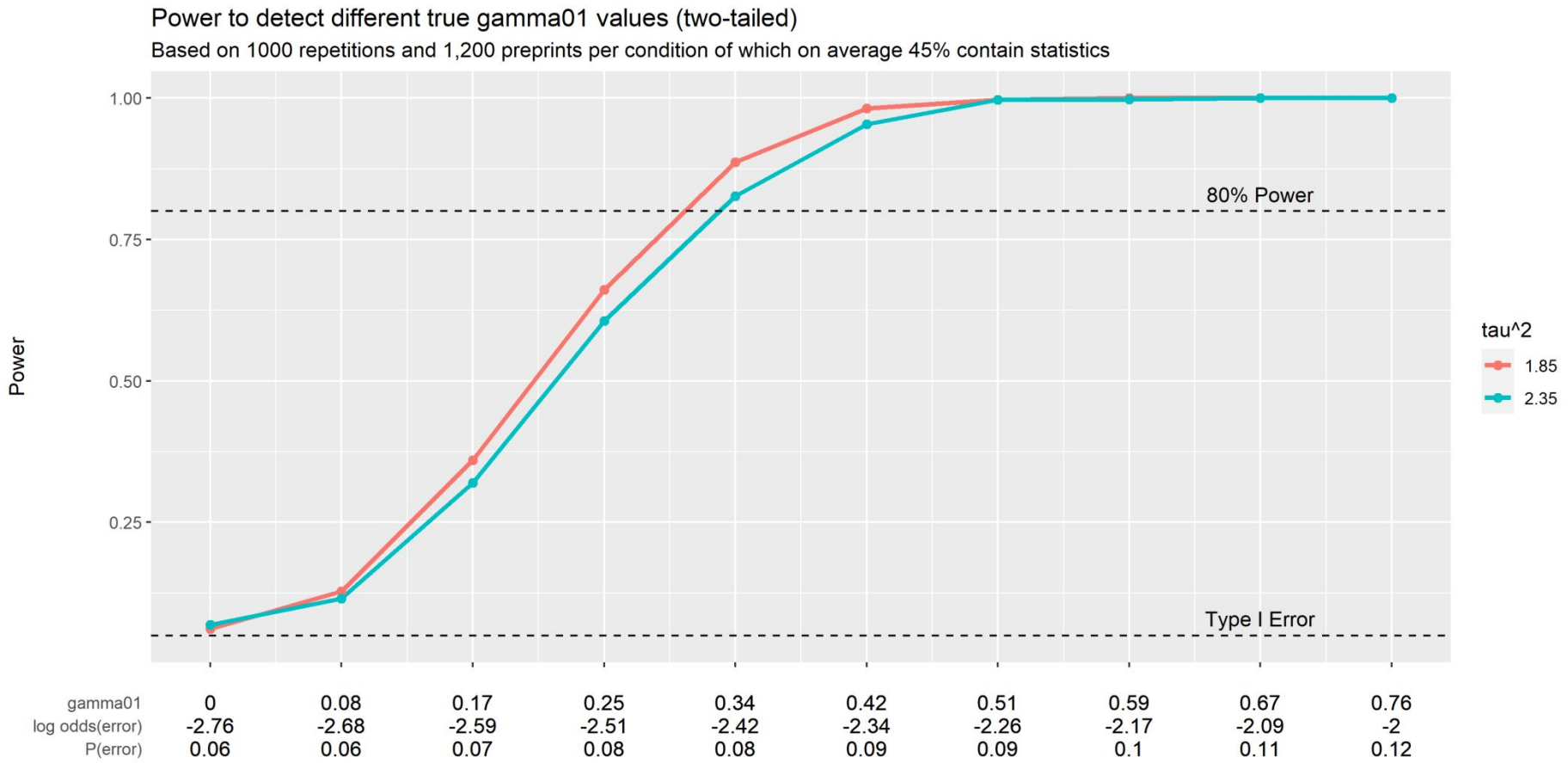

Figure 1. Results of the statistical power analysis.

References

Alexander, P. E., Debono, V. B., Mammen, M. J., Iorio, A., Aryal, K., Deng, D., Brocard, E., & Alhazzani, W. (2020). COVID-19 coronavirus research has overall low methodological quality thus far: Case in point for chloroquine/hydroxychloroquine. *Journal of Clinical Epidemiology*, *123*, 120–126. https://doi.org/10.1016/j.jclinepi.2020.04.016

Anderson, M. S., Ronning, E. A., DeVries, R., & Martinson, B. C. (2010). Extending the Mertonian Norms: Scientists' Subscription to Norms of Research. *The Journal of Higher Education*, *81*(3), 366–393. https://doi.org/10.1353/jhe.0.0095

Bakker, M., Dijk, A. van, & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science: *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691612459060

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678. https://doi.org/10.3758/s13428-011-0089-5

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533. https://doi.org/10.1038/483531a

Brainard, J. (2020, May 13). Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? *Science*. https://www.sciencemag.org/news/2020/05/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat. Accessed 01-02-2021.

Bramstedt, K. A. (2020). The carnage of substandard research during the COVID-19 pandemic: A call for quality. *Journal of Medical Ethics*, *46(12)*, 803–807. https://doi.org/10.1136/medethics-2020-106494

Brown, N. J. L., & Heathers, J. A. J. (2017). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, *8*(4), 363–369. https://doi.org/10.1177/1948550616673876

Callaway, E. (2020). Will the pandemic permanently alter scientific publishing? *Nature*, *582(7811)*, 167–168. https://doi.org/10.1038/d41586-020-01520-4

Carneiro, C. F. D., Queiroz, V. G. S., Moulin, T. C., Carvalho, C. A. M., Haas, C. B., Rayêe, D., Henshall, D. E., De-Souza, E. A., Amorim, F. E., Boos, F. Z., Guercio, G. D., Costa, I. R., Hajdu, K. L., van Egmond, L., Modrák, M., Tan, P. B., Abdill, R. J., Burgess, S. J., Guerra, S. F. S., … Amaral, O. B. (2020). Comparing quality of reporting between preprints and peer-reviewed articles in the biomedical literature. *Research Integrity and Peer Review*, *5(1)*, 16. https://doi.org/10.1186/s41073-020-00101-3

Chiu, W.-T., Huang, J.-S., & Ho, Y.-S. (2004). Bibliometric analysis of severe acute respiratory syndrome-related research in the beginning stage. *Scientometrics*, *61(1)*, 69–77. https://doi.org/10.1023/B:SCIE.0000037363.49623.28

Dinis-Oliveira, R. J. (2020). COVID-19 research: Pandemic versus "paperdemic", integrity, values and risks of the "speed science." *Forensic Sciences Research*, *5(2)*, 174–187. https://doi.org/10.1080/20961790.2020.1767754

Dobler, C. C. (2020). Poor quality research and clinical practice during COVID-19. *Breathe*, *16(2)*. https://doi.org/10.1183/20734735.0112-2020

Epskamp, S., & Nuijten, M. B. (2014). *statcheck: Extract statistics from articles and recompute p-values.* R package version 1.4.0. Retrieved from https://cran.r-project.org/web/packages/statcheck/index.html.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Franco, A., Malhotra, N., & Simonovits, G. (2015). Underreporting in political science survey experiments: Comparing questionnaires to published results. *Political Analysis*, *23(2)*, 306–312.

Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological and Personality Science*, *7*(1), 8–12. https://doi.org/10.1177/1948550615598377

Georgescu, C., & Wren, J. D. (2018). Algorithmic identification of discrepancies between published ratios and

their reported confidence intervals and P-values. *Bioinformatics*, *34(10)*, 1758–1766. https://doi.org/10.1093/bioinformatics/btx811

Gu, X., Mulder, J., & Hoijtink, H. (2017). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *71*, 229–261. https://doi.org/10.1111/bmsp.12110

Higgins J. P. T., Altman D. G., Gøtzsche P. C., Jüni P., Moher D., Oxman A. D., Savovic, J., Schulz, K. F., Weeks, L., & Sterne, J. A. C. (2011) The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal, 343*, d5928. https://doi.org/10.1136/bmj.d5928

Horbach, S. P. J. M. (2020). *Pandemic publishing: Medical journals drastically speed up their publication process for COVID-19.* BioRxiv. https://doi.org/10.1101/2020.04.18.045963

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Jung, R. G., Di Santo, P., Clifford, C., Prosperi-Porta, G., Skanes, S., Hung, A., Parlow, S., Visintini, S., Ramirez, F. D., & Simard, T. (2020). *Methodological rigor in COVID-19 clinical research: A systematic review and case-control analysis.* MedRxiv. https://doi.org/10.1101/2020.07.02.20145102

Marcus, A., & Oransky, I. (n.d.). The science of this pandemic is moving at dangerous speeds. Wired. Retrieved February 1, 2021, from https://www.wired.com/story/the-science-of-this-pandemic-is-moving-at-dangerous-speeds/Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Mulder, J., Williams, D. R., Gu, X., Olsson-Collentine, A., Tomarken, A., Böing-Messing, F., Hoijtink, H., . . . van Lissa, C. (*in press*). BFpack: Flexible Bayes factor testing of scientific theories in R. *Journal of Statistical Software*.

Nuijten, M. B., Borghuis, J., Veldkamp, C. L., Dominguez-Alvarez, L., Van Assen, M. A., & Wicherts, J. M. (2017). Journal data sharing policies and statistical reporting inconsistencies in psychology. *Collabra: Psychology*, *3*(1). https://doi.org/10.1525/collabra.102

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Petrocelli, J. V., Clarkson, J. J., Whitmire, M. B., & Moon, P. E. (2013). When ab ≠ c – c': Published errors in the reports of single-mediator models. *Behavior Research Methods*, *45*(2), 595–601. https://doi.org/10.3758/s13428-012-0262-5

Peyrin-Biroulet, L. (2020). Will the quality of research remain the same during the COVID-19 pandemic? *Clinical Gastroenterology and Hepatology*, *18(9)*, 2142. https://doi.org/10.1016/j.cgh.2020.03.054

R Core Team. (2020). *R: A language and environment for statistical computing.*

Ramanan, M., Stolz, A., Rooplalsingh, R., Billot, L., Myburgh, J., & Venkatesh, B. (2020). An evaluation of the quality and impact of the global research response to the COVID-19 pandemic. *The Medical Journal of Australia*. https://doi.org/10.5694/mja2.50790

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Sweileh, W. M. (2019). A bibliometric analysis of health-related literature on natural disasters from 1900 to 2017. *Health Research Policy and Systems*, *17(1)*, 18. https://doi.org/10.1186/s12961-019-0418-1

Weissgerber, T., Riedel, N., Kilicoglu, H., Labbé, C., Eckmann, P., Riet, G. ter, Byrne, J., Cabanac, G., Capes-Davis, A., Favier, B., Saladi, S., Grabitz, P., Bannach-Brown, A., Schulz, R., McCann, S., Bernard, R., & Bandrowski, A. (2021). Automated screening of COVID-19 preprints: Can we help authors to improve transparency and reproducibility? *Nature Medicine*, *27(1)*, 6–7. https://doi.org/10.1038/s41591-020-01203-7

Wells, G. A., Shea, B, O'Connel, D., Peterson, J, Welch, V., Losos, M., & Tugwell, P. (2021). The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Retrieved from http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp