Scientometrics (2021) 126:9133–9145
https://doi.org/10.1007/s11192-021-04150-3

Current use of effect size or confidence interval analyses in clinical and biomedical research

Emilyane de Oliveira Santana Amaral[1] · **Sergio Roberto Peres Line**[1]

Received: 16 March 2021 / Accepted: 1 September 2021 / Published online: 18 September 2021

Abstract
The isolated use of the statistical hypothesis testing for two group comparison has limitations, and its combination with effect size or confidence interval analysis as complementary statistical tests is recommended. In the present work, we estimate the use of these complementary statistical tests (i.e. effect size or confidence interval) in recently published in research articles in clinical and biomedical areas. *Methods*: The ProQuest database was used to search published studies in academic journals between 2019 and 2020. The analysis was carried out using terms that represent five areas of clinical and biomedical research: "brain", "liver", "heart", "dental", and "covid-19". A total of 119,558 published articles were retrieved. *Results*: The relative use of complementary statistical tests in clinical and biomedical publications was low. The highest frequency usage of complementary statistical tests was among articles that also used statistical hypothesis testing for two-sample comparison. Publications with the term "covid-19" showed the lowest usage rate of complementary statistical tests when all article were analyzed but presented the highest rate among articles that used hypothesis testing. *Conclusion*: The low use of effect size or confidence interval in two-sample comparison suggests that coordinate measures should be taken in order to increase the use of this analysis in clinical and biomedical research. Their use should be emphasized in statistical disciplines for college and graduate students, become a routine procedure in research laboratories, and recommended by reviewers and editors of scientific journals.

Keywords Data Analysis · Methodology · Statistical Analysis · Effect Size

JEL Classification I10 (Health) · c12 (Hypothesis Testing: General) · c18 (Methodological Issues: General)

Mathematics Subject Classification 62–11 (research data for problems pertaining to statistics) · 62P10 (applications of statistics to biology and medical sciences) · 92B15 (general biostatistics)

✉ Sergio Roberto Peres Line
serglin@unicamp.br

Emilyane de Oliveira Santana Amaral
emilyaneoliveiras@gmail.com

[1] Piracicaba Dental School, University of Campinas, Piracicaba, SP, Brazil

Introduction

Statistical analysis comparing two groups or populations is a common procedure in clinical and biomedical research analysis. It has an essential role in data analysis as it can not only allow to test hypothesis but also to measure the strength of the relationship between two groups, allowing a coherent association of the obtained data, leading to adequate conclusions and practical significance (Kraemer, 2014). However, it is important to remember that the results of the statistical inference method assume that the data have been correctly collected, analyzed, and reported, and that they are supported by a statistical model that takes in account the data variation (Greenland et al., 2016), since traditional significance tests do not adequately consider systematic error (Schuemie et al., 2014). Failure to meeting these parameters may compromise the interpretation of the results (Greenland et al., 2016). Due to the relative complexity of the experimental design and the statistical analysis, the reader's confidence in published scientific research frequently relies on the expertise of reviewers' analysis and the authors' conclusion statements that may have different levels of transparency. An extremely summarized representation or, even, the total lack of transparency of the statistical model, creates difficulties for the understanding and critical analysis of the subsequent assumptions (Greenland et al., 2016; Kraemer, 2014). In this context, inadequate use or interpretation of statistical tests can lead to flawed conclusions, and failures in the reproducibility of scientific studies (Kraemer, 2014). These problems may go unnoticed by readers who are not familiar with statistical methods of data analysis, which include the lay public as well as health students and professionals (Jenny et al., 2018).

Widely used in inferential statistical analysis, the p-value is the most frequent parameter used to assess the testing hypothesis. The p-value generated after the hypothesis test is defined as the probability that the chosen test statistic is at least as large as its observed value—if all the model's assumptions are correct (Gigerenzer, 2018; Greenland et al., 2016). However, the real meaning and interpretations of the p-value are accompanied by numerous errors and misinterpretations that have been largely widespread (Gigerenzer, 2018). The p-value should be interpreted with caution, as low scores (usually <0.05) do not imply the probability that a hypothesis is true or replicable, and do not relate to an actual presentation of the differences between the two groups, since it is influenced by the sample size and is not able to provide the magnitude of the effect or its accuracy (Berben et al., 2012; Gigerenzer, 2018; Nakagawa & Cuthill, 2007; Wasserstein & Lazar, 2016). While calculating the p-value, standardized significance levels are established, ignoring the selected sample size and the particularities of each study. Consequently, the p-value is frequently interpreted without considering the real practical clinical or biological meaning, generating absolute and arbitrary conclusions about what is "true" or "false", rather than expressing the level of compatibility between the sample and the hypotheses (Altman & Krzywinski, 2016, 2017; Gigerenzer, 2018; Greenland et al., 2016; Wasserstein & Lazar, 2016). Besides, it is worth mentioning that the p-value is asymmetric, arbitrary, and variable between samples (Altman & Krzywinski, 2017; Espirito Santo & Daniel, 2015). Therefore, the study's analysis, presentation, and conclusions should rely on the combined use of p-value with other statistical tests (Altman & Krzywinski, 2017).

In addition to the debate over the isolated use of the p-value and its misinterpretation, there is concern about the consequent impact generated by the "significance pursuit" (Amrhein et al., 2017). The "data dredging" or "p-hacking" is a bias characterized by carrying out alternative analyses until a significant result is found. This result is reported selectively without the

presence of previously found non-significant results (Amrhein et al., 2017; Bruns & Ioannidis, 2016). Publication and selective reporting biases are part of p-hacking (Amrhein et al., 2017; Chan et al., 2014; Chavalarias et al., 2016; Cristea & Ioannidis, 2018; DeVito et al., 2020; Fanelli, 2012; Goodman, 2019; Lane et al., 2016; Lynch et al., 2007; Rosenthal, 1979; Simmons et al., 2011; Song et al., 2010). These phenomena cause systematic errors in the results and bias the successor statistical tests to the hypothesis test and, consequently, compromise the research results (Amrhein et al., 2017; Bruns & Ioannidis, 2016). In addition, there are so many "researcher degrees of freedom" that, in some cases, researchers choose statistical tests based on the data obtained and not on previous planning. These facts contribute to the generation of biased results and highlight the importance of all data processing steps (Gelman & Loken, 2013; Loken & Gelman, 2017).

The limitations and misunderstandings caused by the isolated p-value report can be overcome by complementing this analysis with other statistics. Confidence intervals and effect size tests have been pointed as valuable alternatives (Baguley, 2009; Cumming, 2014; Durlak, 2009). These complementary tests (i.e. effect size and confidence interval) can indicate the magnitude and practical importance of the results, showing the distance between the actual situation and the null hypothesis (Kraemer, 2014; Nuzzo, 2014). As a p value cannot provide the magnitude and practical importance of an effect, a small p-value can be related to a low, medium, or high effect (Durlak, 2009). Effect size calculations do not depend on the sample size, reduce p-value misinterpretation, and are relevant criteria for comparing results during the elaboration of meta-analysis studies (Espirito Santo & Daniel, 2015; Kraemer, 2014).

Using Bayesian statistical theory, Johnson (2013) estimated that between 17 and 25% of the scientific studies' results reporting borderline significant p-value may be fallacious, compromising their reproducibility. Reproducibility is defined as the ability to duplicate results using the same materials used in the original research (Goodman et al., 2016), and there are ways to estimate the reproducibility rate without performing extensive replications and rigorous methodology. However, it should be noted that there is no consensus to assess replication success, so studies can be examined from a variety of perspectives (Collaboration, 2015; Gelman & Carlin, 2014; Goodman et al, 2016). It is noteworthy that even research of exemplary quality can have empirical results that are irreproducible due to random or systematic errors, the presence of selective reports and analyses, and insufficient specifications of the necessary conditions to obtain the results (Collaboration, 2015; Goodman et al., 2016; Greenland, 2017). Therefore, in the last years, publishing guidelines are increasingly requesting the use of p-value together with effect sizes and/or confidence intervals (Altman & Krzywinski, 2016, 2017; Kraemer, 2014).

The use of effect size and confidence interval in statistical analysis of research articles has traditionally been low (Fidler et al., 2004a, b; Freire et al., 2019; Fritz et al., 2012a, b; Stang et al., 2017), and there is no precise estimate of the current use of these tests in clinical and biomedical research areas. Considering the expressive concern about scientific studies' quality in terms of not using the p-value and effect sizes to generate scientific conclusions, this study aims to analyze and quantify the use of complementary statistical tests (i.e. effect size or confidence interval) in recently published studies in different health areas.

Material and methods

The present study used the ProQuest database (www.proquest.com) to search by the number of published studies in academic journals between 2019 and 2020 considering pre-defined health knowledge areas to analyse and quantify the use of effect size in clinical and biomedical research studies.

The search included studies with the terms "treatment" and one of the following terms: "liver", "brain", "heart", "covid-19", or "dental". The term "treatment" was included to enrich our sample with articles with clinical and biomedical research studies. The anatomical and disease related terms were included to analyse the use of effect size in distinct areas of clinical and biomedical research. In order to select only research articles, papers including the terms "meta-analysis" and "review" were excluded from the analyses.

Association between the number of published studies and the use of complementary statistical tests (i.e. effect size or confidence interval)

Initially, this study sought to establish a relation between the number of published studies and the effect size or confidence interval (herein termed as complementary statistical tests). Therefore, the search results for each area mentioned above were accounted in two different groups:

Group I: Total of published studies in the area.
Group II: Total of published studies in the area that have the terms "effect size", "cohen's d", "hedges' g" or "confidence interval". The terms "cohen's d" or "hedges' g" were included as they are popular effect size methods used in the literature and can be mentioned without the use of the term "effect size".

At this stage of the search strategy, were used combinations and variations of the following terms for each interest area: ("liver/brain/dental/covid-19/heart" AND "treatment"), AND ("effect size" OR "cohen's d" OR "hedges' g" OR "confidence interval").

The combined use of statistical hypothesis and complementary statistical tests

The subsequent stage of this study sought to establish a relation between the number of published studies that used statistical hypothesis testing inference and studies that used complementary statistical tests. In addition to the previously collected data, two other groups were accounted for:

Group III: Total of published studies that have the terms "t-test" or "wilcoxon" or "mann–whitney". These terms were used as they represent frequently used methods for statistical hypothesis testing for groups comparison.
Group IV: Total of published studies that have the terms "t-test" or "wilcoxon" or "mann–whitney" and "effect size" or "cohen's d" or "hedges' g" or "confidence interval".

At this stage of the search strategy, were used combinations and variations of the following terms for each interest area: ("liver/brain/dental/covid-19/heart" AND "treatment"), AND ("effect size" OR "cohen's d" OR "hedges' g" OR "confidence interval"), AND ("t-test" OR "wilcoxon" OR "mann–whitney").

Statistical analysis

The statistical analysis was made by using the R 4.0.2 software (https://www.r-project.org) and the obtained data were analyzed by the Chi-square test with a significance level of 5%. The Cramér's V was used for the effect size analysis and for its interpretations were considered the benchmarks suggested by Cohen (1988): for chi-square tests with degrees of freedom equal to 4, a value of Cramér's V between 0.05 and 0.15 indicates a small effect, a value within the range of 0.15–0.25 indicates a medium effect and a value greater than 0.25 indicates a large effect. Pair-to-pair comparisons were performed between groups using the Test of Equal or Given Proportions with a significance level of 5% and a 95% confidence interval (for Test of Equal or Given Proportions results, see Online Resource).

Results

Effect size and confidence interval analysis are not frequently used in clinical and biomedical research

The results showed that published articles in clinical and biomedical areas make low use of the complementary statistical tests. The search results with the terms "brain", and "liver" have the highest number of articles published (Table 1). The relative use (frequency percent) of complementary tests (Group II/Group I) was restricted to 8.9% with "heart" and "treatment", 7.9% with "liver" and "treatment", 7.3% with the terms "brain" and "treatment", 6.1% with "dental" and "treatment", and 3.3% with "covid-19" and "treatment" (Fig. 1). Among the study areas, there is a statistically perceptible difference with a significance level of 5% in the relative use of effect size ($X^2_4 = 330.24$; p value < 2.2e-16; Cramer's $V = 0.05$). In the Test of Equal or Given Proportions results analysis, it was observed that this difference is present among all groups ($p < 0.01$ in all comparisons).

Table 1 Number of papers according to studied areas and groups

	Group I	Group II	Group III	Group IV
Liver	33,977	2678	11,830	1531
Brain	35,387	2600	11,675	1446
Heart	33,277	2965	9244	1535
Covid-19	8107	267	426	90
Dental	8810	541	2019	237

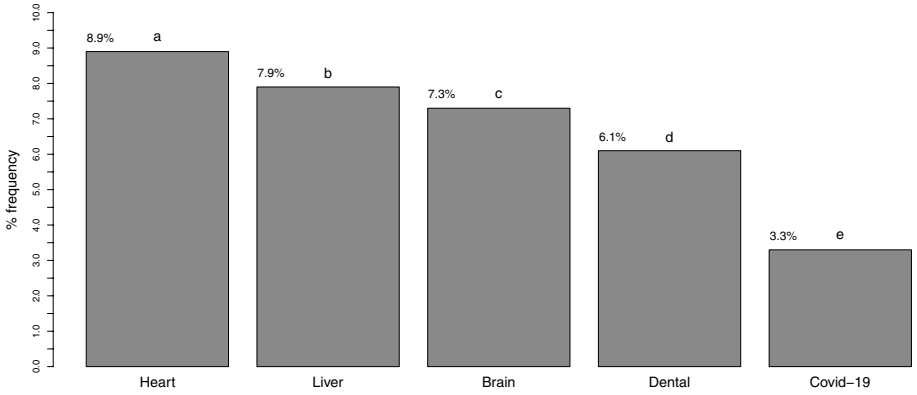

Fig. 1 Frequency of use of complementary tests among all articles in the studied areas

Studies that use statistical hypothesis testing for sample comparison and complementary statistical tests represent the smallest portion of published articles

It is possible that the use of complementary statistical tests may be more frequent in articles that also use statistical hypothesis testing for comparison of two groups. Therefore, we seek to estimate the use of statistical hypothesis testing for comparison of two groups among all published articles. The percent frequency of the articles that used statistical hypothesis testing for group comparison (group III/group I) was 34.8% with the terms "liver" and "treatment", 33% with "brain" and "treatment", 27.8% with "heart" and "treatment", 22.9% with "dental" and "treatment", and 5.3% with "covid-19" and "treatment" (Fig. 2). Among the study areas, there is a statistically notable difference with a significance level of 5% in the use of hypothesis test ($X^2_4 = 3195.6$; p value $< 2.2e\text{-}16$; Cramer's $V = 0.16$). In the Test of Equal or Given Proportions results analysis, it was observed that

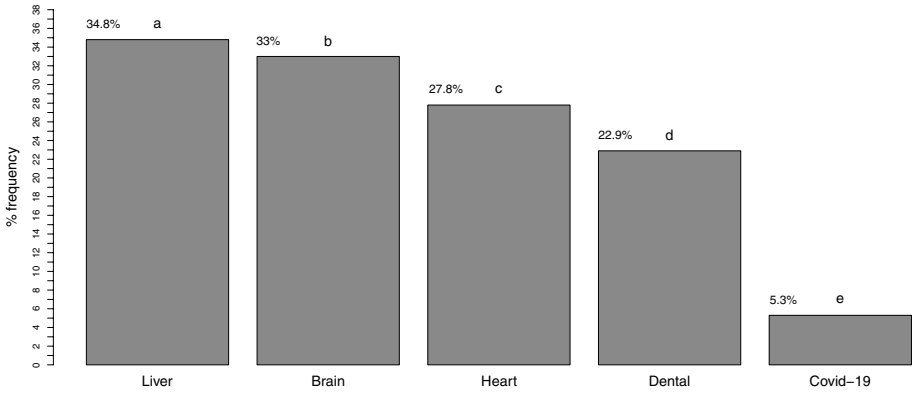

Fig. 2 Frequency of articles that used statistical hypothesis testing for group comparison among all articles in the studied areas

this difference is present among all groups ($p < 0.0001$ in all comparisons). These results indicate that most articles used other methods than two group comparison hypothesis testing. These works may have used multiple groups comparison analysis methods or other types of analysis such as correlation and regression, Bayesian methods, or presented only a descriptive analysis.

The relative frequency of articles that used both statistical hypothesis and complementary testing for two sample comparisons among all articles (Group IV/Group I) was 4.6% with "heart" and "treatment", 4.5% with "liver" and "treatment", 4.1% with the terms "brain" and "treatment", 2.7% with "dental" and "treatment", and 1.1% with "covid-19" and "treatment" (Fig. 3). Among the study areas, there is a statistically notable difference with a significance level of 5% in the articles that used hypothesis test and complementary statistical tests among all articles ($X^2_4 = 267.82$; p value < 2.2e-16; Cramer's $V = 0.05$). In the Test of Equal or Given Proportions results analysis, it was observed that there was no statistically perceptible difference between "liver" and "heart" (p value = 0.52). While "heart", "dental" and "covid-19" showed a notable statistical difference among all areas ($p < 0.01$ in all comparisons, Fig. 3).

Among articles that used statistical hypothesis testing for two sample comparison the use of complementary statistical tests (Group IV/Group III) was 21.1% with "covid-19" and "treatment", 16.6% with "heart" and "treatment", 12.9% with "liver" and "treatment", 12.4% with the terms "brain" and "treatment", and 11.7% with "dental" and "treatment" (Fig. 4). Among the study areas, there is a statistically perceptible difference with a significance level of 5% in the use of complementary statistical tests among articles that use hypothesis test ($X^2_4 = 114.84$; p value < 2.2e-16; Cramer's $V = 0.06$). In the results analysis of the Test of Equal or Given Proportions, it was observed that there was no statistically notable difference between "liver", "brain" and "dental" (p value = 0.21, 0.14, and 0.43). While the "heart" and "covid-19" areas showed a notable statistical difference among all areas ($p < 0.02$ in all comparisons, Fig. 4).

It is possible that some articles of Group IV could have only mentioned terms related to statistical hypothesis testing ("t-test", "wilcoxon" or "mann–whitney") and the complementary statistical tests ("effect size", "cohen's d" or "hedges' g" or "confidence interval")

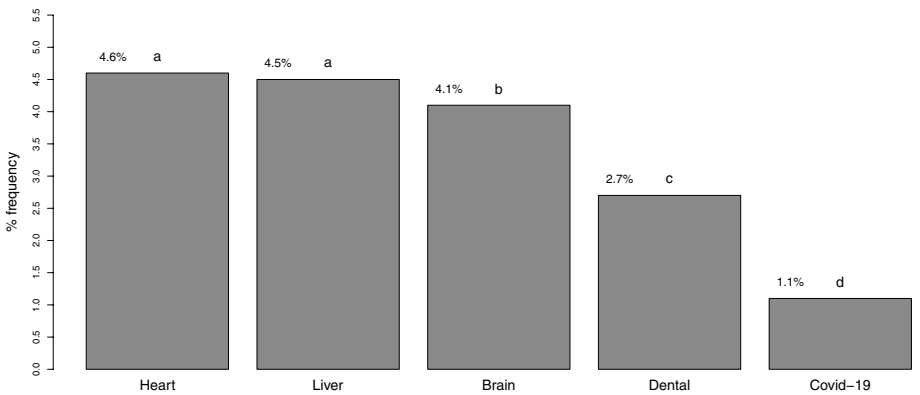

Fig. 3 Frequency of articles that used both statistical hypothesis and complementary testing for two sample comparisons among all articles in the studied areas

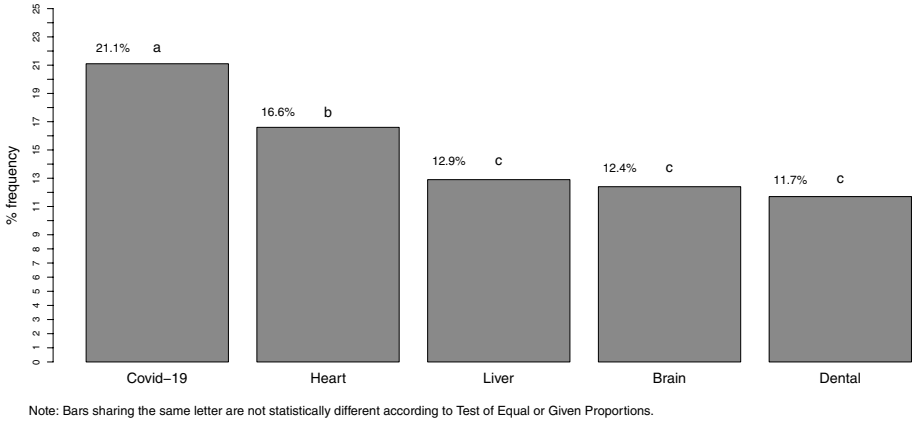

Fig. 4 Frequency of complementary statistical tests among articles that used statistical hypothesis testing for two sample comparison in the studied areas

without having performed analysis with these tests. In order to have an appraisal of the proportion of articles in Group IV that have performed both tests, we inspected a sample of 125 articles randomly selected. One hundred and nineteen out of the 125 papers analyzed (95.2% of the papers) had used both tests in their data analysis, and the 6 remaining articles (4.8%) had used only hypothesis testing. These results show that the combined use of statistical hypothesis testing and complementary tests is likely to be slightly lower than our initial prediction.

Discussion

The use of *p* value in isolation has been the subject of discussions and has raised concerns about the replicability, veracity, and reliability of the conclusions generated (Peng, 2015; Wasserstein & Lazar, 2016). In agreement with the literature (Berben et al., 2012), this study shows that estimates of effect size are generally not reported as part of the scientific studies results.

Previous studies focused on eating and psychological disorders (Crosby et al., 2006), psychology (Faulkner et al., 2008; Fidler et al, 2005; Fritz et al., 2012a, b), educational psychology (Osborne, 2008; Sun et al., 2010), and learning and education (Barry et al., 2016) are in consonance with the results presented in this research since it was observed that the effect size analysis is not routinely used in clinical and biomedical research. However, it is important to point out that the present study was performed without applying filters related to the magazine type, impact factor, or revision criteria, such as mandatory complementary statistical test report or peer reviews. It is noteworthy that the complementary statistical test report analysis might vary according to the used analyzes, criteria established by magazines for publishing the articles and the type filter applied to select the articles (Alhija & Levy, 2009; Sun et al., 2010). The authors' lack of knowledge and the incorrect use and interpretation of statistical analysis (Peng, 2015) contribute to this context, in which complementary statistical tests are not used. It is worth mentioning that studies that resulted in a statistically non-significant p-value should present the results of their tests (American Psychological Association, 2010) and, even if this is the recommendation,

the authors tend not to mention the complementary statistical tests in these cases (Berben et al., 2012).

The small prevalence of the articles that use both hypothesis testing and complementary statistical tests can be justified by the low usage of complementary tests. Factors that contribute to this aspect include the entrenched habit of using the p value as the main analytical and determinant method, the lack of presentation and interpretation of confidence intervals that indicate the direction and size of the treatment effect, and even the presence of "spin" (incorrect and selective representation, or misrepresentation of search results) (Gates & Ealing, 2019; Gewandter et al., 2015, 2017). Standardized significance levels are the most demanded parameters by the academic community and journals and are also highly emphasized in academic courses in statistics (Wasserstein & Lazar, 2016). In addition to the above-mentioned reasons, the null hypothesis test establishes an arbitrary p-value and creates a belief that "significative" discoveries (p value < 0.05) are more valuable, reliable, and reproducible (Nickerson, 2000) and that results with p values higher than 0.05 are not relevant (Ialongo, 2016). This reality makes it more likely to publish studies with "positive" than with "negative" results and erroneously reaffirm the hypothesis test as the main statistical method (Begg & Berlin, 1988). In a retrospective analysis of 136,212 clinical trials performed between 1975 and 2014, it was found that the statistical power increased (although they are still small in most cases—around 10% average power) while the use of effect size remained stable. This increase was mainly due to increasing sample sizes (Lamberink et al., 2018). The widespread use of p value to generate scientific conclusions tends to be predominant in articles in the most diverse areas (Fidler et al., 2004a, b; Kirk, 2001), holding a dominant position in the statistical analyses to obtain conclusions (Fidler et al., 2004a, b). The null hypothesis significance test favors the aggravation of cognitive distortions (Greenland, 2017), that remain due to the researchers' internalized belief in the "null ritual" and the desire to obtain a "significant p-value" (Gigerenzer, 2018; Meehl, 1978), although the "significant differences" found are little more than complex and causally uninterpretable results of statistical power functions (Meehl, 1978). Based on this, with sufficient data, it is possible to reject any null hypothesis, and after the study has been completed, it may be possible that the alternative hypothesis becomes the desirable alternative hypothesis (Gelman, 2016; Yarkoni, 2020).

Statistical methods do not free data from uncertainty and that they provide a noisy signal (Greenland, 2017; Wasserstein et al., 2019). It should always be considered that every scientific dataset comes with its own systematic errors skewing the observed distributions away from the null (Gelman & Carlin, 2017; Gigerenzer, 2018). In order to maximize the chances of producing reliable and meaningful data, the design and execution of a research work should be carefully planned before its execution (Wasserstein et al., 2019). The concomitant use of the p value, effect size analysis, and/or confidence intervals would allow more precise and reliable conclusions (Altman & Krzywinski, 2017; Wasserstein et al., 2019). It is also desirable that studies have high or at least reasonable statistical power (Gigerenzer, 2018). The author-reader communication must be clear and transparent about the confidence level present in the results obtained in the statistical analysis (Greenland, 2017; Wasserstein et al., 2019).

It is worth mentioning that the "covid-19" articles have the lowest use of statistical hypothesis testing for sample comparison, complementary statistical, and the lowest concomitant use of these two methods among all articles published in the area. These results agree with other published articles reporting a lower scientific quality and accuracy (Zdravkovic et al., 2020), and a high rate of post-publication corrections and retractions (Soltani & Patini, 2020). These may occur as a side effect of rushed publications induced

by the lack of information and sudden high interest in this subject matter. However, among the “covid-19" papers that have used hypothesis testing the use of complementary statistical tests is significantly higher than in the other areas. Our results suggest that careful analysis of the published data, tables, and figures should be made by the reader before complying with the authors’ claims in papers related to covid-19.

While the statistical hypothesis test provides the p-value that represents a statistical summary of the compatibility between the observed data and what we would expect to observe (Greenland et al., 2016), the effect size is represented by a number that measures the strength of the relationship between the groups. These tests used in conjunction will give a more accurate appraisal of the relationship between two sample populations. The estimation and interpretation of effect size are straightforward and can be made online using diverse reliable sources (Becker, 2000; Lenhard & Lenhard, 2016). The results of this article suggest that coordinate measures should be taken to increase the use of effect size in research analysis. Its use should be emphasized in statistical disciplines for college and graduate students, become a routine procedure in research laboratories, and recommended by reviewers and editors of scientific journals.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11192-021-04150-3.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by EOSA and SRPL The first draft of the manuscript was written by EOSA and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding EOSA was supported by a grant from CAPES #8887.513474/2020–00, SRPL was supported by Grant from Brazilian Scientific Council CNPq # 305783/2018–1.

Data Availability All authors certify that all data and materials support the statements of the article and comply with field standards.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Alhija, F. N. A., & Levy, A. (2009). Effect size reporting practices in published articles. *Educational and Psychological Measurement, 69*(2), 245–265. https://doi.org/10.1177/0013164408315266

Altman, N., & Krzywinski, M. (2016). Points of significance: P values and the search for significance. *Nature Methods, 14*(1), 3–4. https://doi.org/10.1038/nmeth.4120

Altman, N., & Krzywinski, M. (2017). Points of Significance: Interpreting P values. *Nature Methods, 14*(3), 213–214. https://doi.org/10.1038/nmeth.4210

American Psychological Association. (2010). Publication Manual of the American Psychological Association (6.[a] ed.).

Amrhein, V., Korner-Nievergelt, F., & Roth, T. (2017). The earth is flat (p > 0:05): Significance thresholds and the crisis of unreplicable research. *PeerJ*. https://doi.org/10.7717/peerj.3544

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*(3), 603–617. https://doi.org/10.1348/000712608X377117

Barry, A. E., Szucs, L. E., Reyes, J. V., Ji, Q., Wilson, K. L., & Thompson, B. (2016). Failure to report effect sizes: The handling of quantitative results in published health education and behavior research. *Health Education and Behavior, 43*(5), 518–527. https://doi.org/10.1177/1090198116669521

Becker, L. (2000). Effect size Calculators, Effect Size (ES). *University of Colorado Colorado Retrieved from* http://www.uccs.edu/lbecker/effect-size.html*, (1993).*

Begg, C. B., & Berlin, J. A. (1988). Publication Bias : A Problem in Interpreting Medical Data Author (s): Colin B . Begg and Jesse A . Berlin Published by : Wiley for the Royal Statistical Society Stable URL : https://www.jstor.org/stable/2982993, *151*(3), 419–463.

Berben, L., Sereika, S. M., & Engberg, S. (2012). Effect size estimation: Methods and examples. *International Journal of Nursing Studies, 49*(8), 1039–1047. https://doi.org/10.1016/j.ijnurstu.2012.01.015

Bruns, S. B., & Ioannidis, J. P. A. (2016). P-curve and p-hacking in observational research. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0149144

Chan, A. W., Song, F., Vickers, A., Jefferson, T., Dickersin, K., Gøtzsche, P. C., et al. (2014). Increasing value and reducing waste: Addressing inaccessible research. *The Lancet, 383*(9913), 257–266. https://doi.org/10.1016/S0140-6736(13)62296-5

Chavalarias, D., Wallach, J. D., Li, A. H. T., & Ioannidis, J. P. A. (2016). Evolution of reporting P values in the biomedical literature, 1990–2015. *JAMA - Journal of the American Medical Association, 315*(11), 1141–1148. https://doi.org/10.1001/jama.2016.1952

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (Second Edi.). Hisdale: Lawrence Erlbaum Associates.

Collaboration, O. (2015). Estimating the Reproducibility of Psychological Science. *American Association for the Advancement of Science*. https://doi.org/10.31219/osf.io/447b3

Cristea, I. A., & Ioannidis, J. P. A. (2018). P values in display items are ubiquitous and almost invariably significant: A survey of top science journals. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0197440

Crosby, R. D., Wonderlich, S. A., Mitchell, J. E., de Zwaan, M., Engel, S. G., Connolly, K., et al. (2006). An empirical analysis of eating disorders and anxiety disorders publications (1980–2000)–part II: Statistical hypothesis testing. *International Journal of Eating Disorders, 39*(1), 49–54.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science., 25*, 7–29. https://doi.org/10.1177/0956797613504966

DeVito, N. J., Bacon, S., & Goldacre, B. (2020). Compliance with legal requirement to report clinical trial results on ClinicalTrials.gov: a cohort study. *The Lancet, 395*(10221), 361–369. https://doi.org/10.1016/S0140-6736(19)33220-9

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology, 34*(9), 917–928. https://doi.org/10.1093/jpepsy/jsp004

Espirito Santo, H., & Daniel, F. B. (2015). Calcular e apresentar tamanhos do efeito em trabalhos científicos (1): As limitações do p < 0,05 na análise de diferenças de médias de dois grupos. *Revista Portuguesa De Investigação Comportamental e Social, 1*(1), 3–16. https://doi.org/10.7342/ismt.rpics.2015.1.1.14

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Faulkner, C., Fidler, F., & Cumming, G. (2008). The value of RCT evidence depends on the quality of statistical analysis. *Behaviour Research and Therapy, 46*(2), 270–281. https://doi.org/10.1016/j.brat.2007.12.001

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., et al. (2005). Toward improved statistical reporting in the Journal of Consulting and Clinical Psychology. *Journal of Consulting and Clinical Psychology, 73*(1), 136–143. https://doi.org/10.1037/0022-006X.73.1.136

Fidler, F., Geoff, C., Mark, B., & Neil, T. (2004a). Statistical reform in medicine, psychology and ecology. *Journal of Socio-Economics, 33*(5), 615–630. https://doi.org/10.1016/j.socec.2004.09.035

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004b). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science, 15*(2), 119–126.

Freire, A. P. C. F., Elkins, M. R., Ramos, E. M. C., & Moseley, A. M. (2019). Use of 95% confidence intervals in the reporting of between-group differences in randomized controlled trials: Analysis of a representative sample of 200 physical therapy trials. *Brazilian Journal of Physical Therapy, 23*(4), 302–310. https://doi.org/10.1016/j.bjpt.2018.10.004

Fritz, A., Scherndl, T., & Kühberger, A. (2012a). A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough? *Theory & Psychology, 23*(1), 98–122. https://doi.org/10.1177/0959354312436870

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012b). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*(1), 2–18. https://doi.org/10.1037/a0024338

Gates, S., & Ealing, E. (2019). Reporting and interpretation of results from clinical trials that did not claim a treatment difference: Survey of four general medical journals. *BMJ Open*. https://doi.org/10.1136/bmjopen-2018-024785

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem. *Unpublished manuscript*. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gelman, A. (2016). The problems with p-values are not just with p-values. *The American Statistician*, 1–2.

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science, 9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A., & Carlin, J. (2017). Some Natural Solutions to the p-Value Communication Problem—and Why They Won't Work. *Journal of the American Statistical Association, 112*(519), 899–901. https://doi.org/10.1080/01621459.2017.1311263

Gewandter, J. S., Mcdermott, M. P., Kitt, R. A., Chaudari, J., Koch, J. G., Evans, S. R., et al. (2017). Interpretation of CIs in clinical trials with non-significant results: Systematic review and recommendations. *BMJ Open*. https://doi.org/10.1136/bmjopen-2017-017288

Gewandter, J. S., McKeown, A., McDermott, M. P., Dworkin, J. D., Smith, S. M., Gross, R. A., et al. (2015). Data interpretation in analgesic clinical trials with statistically nonsignificant primary analyses: An ACTTION systematic review. *The Journal of Pain, 16*(1), 3–10. https://doi.org/10.1016/j.jpain.2014.10.003

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science, 1*(2), 198–218. https://doi.org/10.1177/2515245918771329

Goodman, S. N. (2019). Why is Getting Rid of P-Values So Hard? Musings on Science and Statistics. *American Statistician, 73*(sup1), 26–30. https://doi.org/10.1080/00031305.2018.1558111

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Getting to Good: Research Integrity in the Biomedical Sciences*. https://doi.org/10.1126/scitranslmed.aaf5027

Greenland, S. (2017). Invited Commentary: The Need for Cognitive Science in Methodology. *American Journal of Epidemiology, 186*(6), 639–645. https://doi.org/10.1093/aje/kwx259

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Ialongo, C. (2016). Understanding the effect size and its measures. *Biochemia Medica, 26*(2), 150–163.

Jenny, M. A., Keller, N., & Gigerenzer, G. (2018). Assessing minimal medical statistical literacy using the Quick Risk Test: A prospective observational study in Germany. *BMJ Open*. https://doi.org/10.1136/bmjopen-2017-020847

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America, 110*(48), 19313–19317. https://doi.org/10.1073/pnas.1313476110

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61*(2), 213–218. https://doi.org/10.1177/00131640121971185

Kraemer, H. C. (2014). Effect Size. *The Encyclopedia of Clinical Psychology*. https://doi.org/10.1002/9781118625392.wbecp048

Lamberink, H. J., Otte, W. M., Sinke, M. R. T., Lakens, D., Glasziou, P. P., Tijdink, J. K., & Vinkers, C. H. (2018). Statistical power of clinical trials increased while effect size remained stable: An empirical analysis of 136,212 clinical trials between 1975 and 2014. *Journal of Clinical Epidemiology, 102*, 123–128. https://doi.org/10.1016/j.jclinepi.2018.06.014

Lane, A., Luminet, O., Nave, G., & Mikolajczak, M. (2016). Is there a Publication Bias in Behavioural Intranasal Oxytocin Research on Humans? Opening the File Drawer of One Laboratory. *Journal of Neuroendocrinology*. https://doi.org/10.1111/jne.12384

Lenhard, W., & Lenhard, A. (2016). Calculation of Effect Sizes. *Psychometrica*. http://www.psychometrica.de/effect_size.html

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Lynch, J. R., Cunningham, M. R. A., Warme, W. J., Schaad, D. C., Wolf, F. M., & Leopold, S. S. (2007). Commercially funded and United States-based research is more likely to be published; good-quality studies with negative outcomes are not. *Journal of Bone and Joint Surgery - Series A, 89*(5), 1010–1018. https://doi.org/10.2106/JBJS.F.01152

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews, 82*(4), 591–605. https://doi.org/10.1111/j.1469-185X.2007.00027.x

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241–301. https://doi.org/10.1037/1082-989X.5.2.241

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature, 506*(7487), 150.

Osborne, J. W. (2008). Sweating the small stuff in educational psychology: How effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. *Educational Psychology, 28*(2), 151–160. https://doi.org/10.1080/01443410701491718

Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance, 12*(3), 30–32. https://doi.org/10.1111/j.1740-9713.2015.00827.x

Rosenthal, R. (1979). The "file drawer" problem and tolerance for null results. *Psychological Bulletin, 86*, 638.

Schuemie, M. J., Ryan, P. B., Dumouchel, W., Suchard, M. A., & Madigan, D. (2014). Interpreting observational studies: Why empirical calibration is needed to correct p-values. *Statistics in Medicine, 33*(2), 209–218. https://doi.org/10.1002/sim.5925

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Soltani, P., & Patini, R. (2020). Retracted COVID-19 articles: A side-effect of the hot race to publication. *Scientometrics, 125*(1), 819–822. https://doi.org/10.1007/s11192-020-03661-9

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., et al. (2010). Dissemination and publication of research findings: An updated review of related biases. *Health Technology Assessment, 14*(8), 1–220. https://doi.org/10.3310/hta14080

Stang, A., Deckert, M., Poole, C., & Rothman, K. J. (2017). Statistical inference in abstracts of major medical and epidemiology journals 1975–2014: A systematic review. *European Journal of Epidemiology, 32*(1), 21–29.

Sun, S., Pan, W., & Wang, L. L. (2010). A Comprehensive Review of Effect Size Reporting and Interpreting Practices in Academic Journals in Education and Psychology. *Journal of Educational Psychology, 102*(4), 989–1004. https://doi.org/10.1037/a0019507

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond $p < 0.05$. *American Statistician, 73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X20001685

Zdravkovic, M., Berger-Estilita, J., Zdravkovic, B., & Berger, D. (2020). Scientific quality of COVID-19 and SARS CoV-2 publications in the highest impact medical journals during the early phase of the pandemic: A case control study. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0241826