Journal of Clinical Epidemiology 102 (2018) 123−128

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Statistical power of clinical trials increased while effect size remained stable: an empirical analysis of 136,212 clinical trials between 1975 and 2014

Herm J. Lamberink[a,*,1], Willem M. Otte[a,b,1], Michel R.T. Sinke[b], Daniël Lakens[c], Paul P. Glasziou[d], Joeri K. Tijdink[e], Christiaan H. Vinkers[f]

[a]*Department of Child Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht and Utrecht University, P.O. Box 85090, Utrecht 3508 AB, The Netherlands*

[b]*Biomedical MR Imaging and Spectroscopy group, Center for Image Sciences, University Medical Center Utrecht and Utrecht University, Heidelberglaan 100, Utrecht 3584 CX, The Netherlands*

[c]*School of Innovation Sciences, Eindhoven University of Technology, Den Dolech 1, Eindhoven 5600 MB, The Netherlands*

[d]*Centre for Research in Evidence-Based Practice, Faculty of Health Sciences and Medicine, Bond University, Gold Coast, Queensland, Australia*

[e]*Department of Philosophy, VU University, De Boelelaan 1105, Amsterdam 1081 HV, The Netherlands*

[f]*Department of Psychiatry, Brain Center Rudolf Magnus, University Medical Center Utrecht and Utrecht University, Heidelberglaan 100, Utrecht 3584 CX, The Netherlands*

Accepted 28 June 2018; Published online 5 July 2018

Abstract

Objectives: To study the statistical power of randomized clinical trials and examine developments over time.

Study Design and Setting: We analyzed the statistical power in 136,212 clinical trials between 1975 and 2014 extracted from meta-analyses from the Cochrane database of systematic reviews. We determined study power to detect standardized effect sizes, where power was based on the meta-analyzed effect size. Average power, effect size, and temporal patterns were examined for all meta-analyses and a subset of significant meta-analyses.

Results: The number of trials with power ≥80% was low (7%) but increased over time: from 5% in 1975−1979 to 9% in 2010−2014. In significant meta-analyses, the proportion of trials with sufficient power increased from 9% to 15% in these years (median power increased from 16% to 23%). This increase was mainly due to increasing sample sizes, while effect sizes remained stable with a median Cohen's h of 0.09 (interquartile range 0.04−0.22) and a median Cohen's d of 0.20 (0.11−0.40).

Conclusion: This study demonstrates that sufficient power in clinical trials is still problematic, although the situation is slowly improving. Our data encourage further efforts to increase statistical power in clinical trials to guarantee rigorous and reproducible evidence-based medicine.

Keywords: Statistical power; Clinical trial; Randomized

1. Introduction

The practice of conducting scientific studies with low statistical power has been consistently criticized across academic disciplines [1−5]. Statistical power is the probability that a study will detect an effect when there is a true effect to be detected. Underpowered studies have a low chance of detecting true effects and have been related to systematic biases including inflated effect sizes and low reproducibility [6,7]. Low statistical power has been demonstrated, among others, in the fields of neuroscience, economics, and psychology [4,8−10]. For clinical trials in the field of medicine, the issue of sample size evaluation

Funding: This work was supported by The Netherlands Organisation for Health Research and Development (ZonMW) grant ''Fostering Responsible Research Practices'' (445001002). The funding source had no involvement in study design; the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Conflict of interest: none.

[1] Authors contributed equally.

* Corresponding author. Department of Child Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht and Utrecht University, Room KC 03.063.0, P.O. Box 85090, Utrecht 3508 AB, The Netherlands. Tel.: +31 88 755 6030; fax: +31 88 755 5350.

E-mail address: h.j.lamberink@umcutrecht.nl (H.J. Lamberink).

https://doi.org/10.1016/j.jclinepi.2018.06.014

What is new?

Key findings

- Study power in clinical trials is low: 7% of trials were sufficiently powered (≥0.8) and 14% had a power above 0.5; within significant meta-analyses 12% was sufficiently powered and 24% had a power above 0.5.
- The percentage of sufficiently powered studies has increased from 5% in 1975−1979 to 9% in 2010−2014.
- Average effect sizes are small and did not increase over time.

What this adds to what was known?

- Trial sizes and study power increased over time, although both are still small in most cases.

What is the implication and what should change now?

- When determining the required sample size of a clinical trial, small effects should be assumed to ensure an adequate sample size.

and statistical power is essential because clinical decision-making and future research are based on these clinical trials [11,12]. Moreover, low power in clinical trials may be unethical in light of the low informational value from the outset while exposing participants to interventions with possible negative (side) effects [1]. Also in medical research, statistical power is low [3,8], but a systematic overview of temporal patterns of power, sample sizes, and effect sizes across medical fields does not exist. In the present study, we provide a comprehensive overview of study power, sample size, and effect size estimates of clinical trials published since 1975, which are included in the Cochrane database of systematic reviews, and analyze emerging trends over time.

2. Materials and methods

Data were extracted and calculated from trials included in published reviews from the second issue of the 2017 Cochrane database of systematic reviews. Cochrane reviews only include meta-analyses if the methodology and outcomes of the included trials are comparable across study populations. Meta-analysis data are available for download in standardized XML-format for those with an institutional Cochrane Library license. We provide open-source software to convert these data and reproduce our entire processing pipeline [13].

Trials were selected if they were published after 1974 and if they were included in a meta-analysis based on at least five trials. Because relatively few studies from 2015 to 2017 were included in our meta-analyses, these years were excluded. For each individual clinical trial, publication year, outcome estimates (odds or risk ratio, risk difference, or standardized mean difference), and group sizes were extracted. For the main analyses, all meta-analyses were used; subanalyses were performed on only the meta-analyses with a reported *P*-value below 0.05, irrespective of the *P*-value of the individual trial. For meta-analyses reporting standardized mean differences (Cohen's d), the reported meta-analytic effect size was used to compute individual study power. For meta-analyses reporting dichotomous outcomes, meta-analytic effect size (Cohen's h) was computed using arcsine transformation of proportions [12]. The main analysis used the effect size extracted from the meta-analysis, which was performed as either fixed or random effects as judged by the authors of that specific Cochrane review. As a sensitivity analysis, we recomputed the meta-analytic effect size using fixed effects, random effects, and unrestricted weighted least squares/weighed average of the adequately powered [14]. This latter method was developed to optimize results from meta-analysis in the context of selective reporting bias: weighted least squares/weighed average of the adequately powered performs better than both fixed and random effects analyses in the context of publication bias, allows to correct for heterogeneity, and gives similar results to fixed effects when both are not present [14,15]. Study power was computed in R using the ''pwr'' package [16]. Following minimum recommendations for the statistical power of studies [12], comparisons with a power above or equal to 80% were considered to be sufficiently powered. Study power, group sizes, and effect sizes over time were summarized and visualized for all clinical trials.

3. Results

Data from 136,212 clinical trials were available, from 11,852 meta-analyses in 1,918 Cochrane reviews. Of these, 77,947 trials (57.2%) were from a meta-analysis with an overall *P*-value below 0.05, from 5,903 meta-analyses (49.8%) in 1,411 Cochrane reviews (73.6%). In the original systematic reviews, fixed effects were used in 55% of meta-analyses, whereas 45% used random effects. Of all trials, 7.3% had a statistical power of at least 80% (the recommended minimum [12], which we shall denote as ''sufficient power'') to detect an effect size as large as the meta-analyzed effect size; for the subset of significant meta-analyses this was 12.4%. The median power (interquartile range [IQR]) was 9% (95% confidence interval [CI] 6%−26%), which was 20% (10%−48%) for significant meta-analysis (Table 1).

Between 1975−1979 and 2010−2014 the proportion of sufficiently powered studies rose from 5.1% (4.3−6.1) to

Table 1. Proportion studies with sufficient power and median power

Selected meta-analyses	N meta-analyses (N studies)	Proportion sufficient (≥0.8) power (95% CI)	Median power (IQR)
All	11,852 (136,212)	7.3 (7.2–7.5)	0.09 (0.06–0.26)
Significant	5,903 (77,947)	12.4 (12.2–12.7)	0.20 (0.10–0.48)

Abbreviations: CI, confidence interval; IQR, interquartile range.

8.5% (8.0–8.9) (Fig. 1, top right), whereas the median power changed from 0.09 (IQR 0.06–0.26) to 0.10 (IQR 0.06–0.30) (Fig. 1, left). Within significant meta-analyses, the rise was more clear: study power increased with the median rising from 16% (IQR 10–39) to 23% (IQR 12–55) (Supplementary Figure 1, left), and the proportion of sufficiently powered studies from 9.0% (7.6–10.6) to 14.7% (13.9–15.5) (Supplementary Figure 1, top right). This trend is seen across medical disciplines (Supplementary Figure 2). When the threshold for sufficient power is set at a minimum of 50% power, the proportion of trials with sufficient power is still low but also rising (Supplementary Figure 3). The distribution of power showed a bimodal pattern, with many low-powered studies and a small peak of studies with power approaching 100% (Fig. 1 and Supplementary Figure 1, bottom right).

The average number of participants enrolled in a trial arm increased over time (Fig. 2, top left). The median group size in 1975–1979 ranged between 30 and 45; for the years 2010–2014 the median group size was between 74 and 92. The median effect sizes are summarized in Table 2; these remained stable over time (Fig. 2). The standardized effect sizes were small, with a median Cohen's h of 0.09 (0.04–0.22) and a median Cohen's d of 0.20 (0.11–0.40) (Table 2); Figure 3 shows the distribution plots for these two measures; for the significant meta-analyses, the median effect sizes were higher (Supplementary Table 1 and Supplementary Figure 4).

Sensitivity analyses showed robust results regardless of the method for performing meta-analysis. The proportion of studies with sufficient power was between 7.2% and 7.5% depending on the method; the median power remained 9% across methods (Supplementary Table 2).

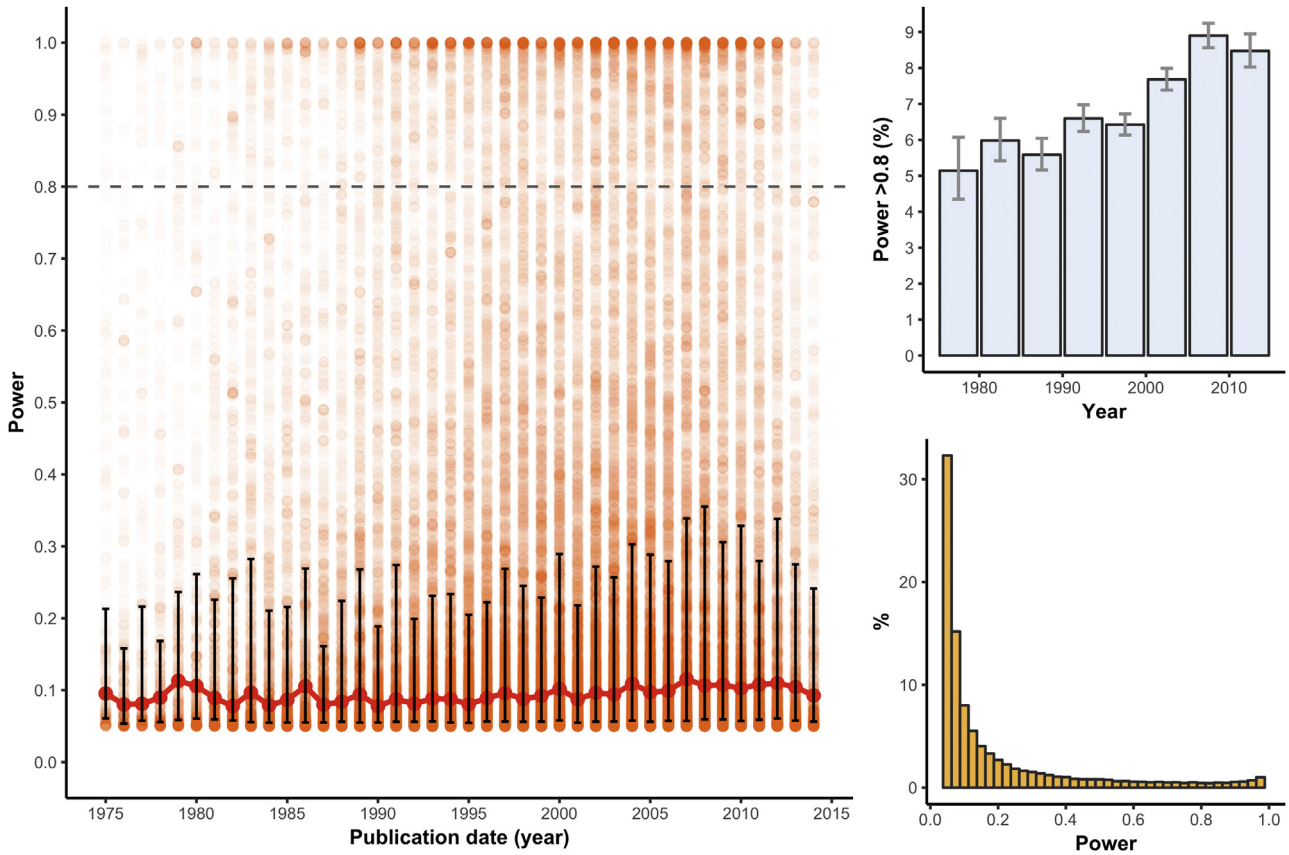

Fig. 1. Statistical power of clinical trials between 1975 and 2014 (left). Individual comparisons are shown as semitransparent dots. Median power is shown in red with interquartile range as error bars. The percentage of adequately powered trial comparisons (i.e., ≥80% power) is increasing over time (top right). The biphasic power distribution of the trials in general is apparent (bottom right). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

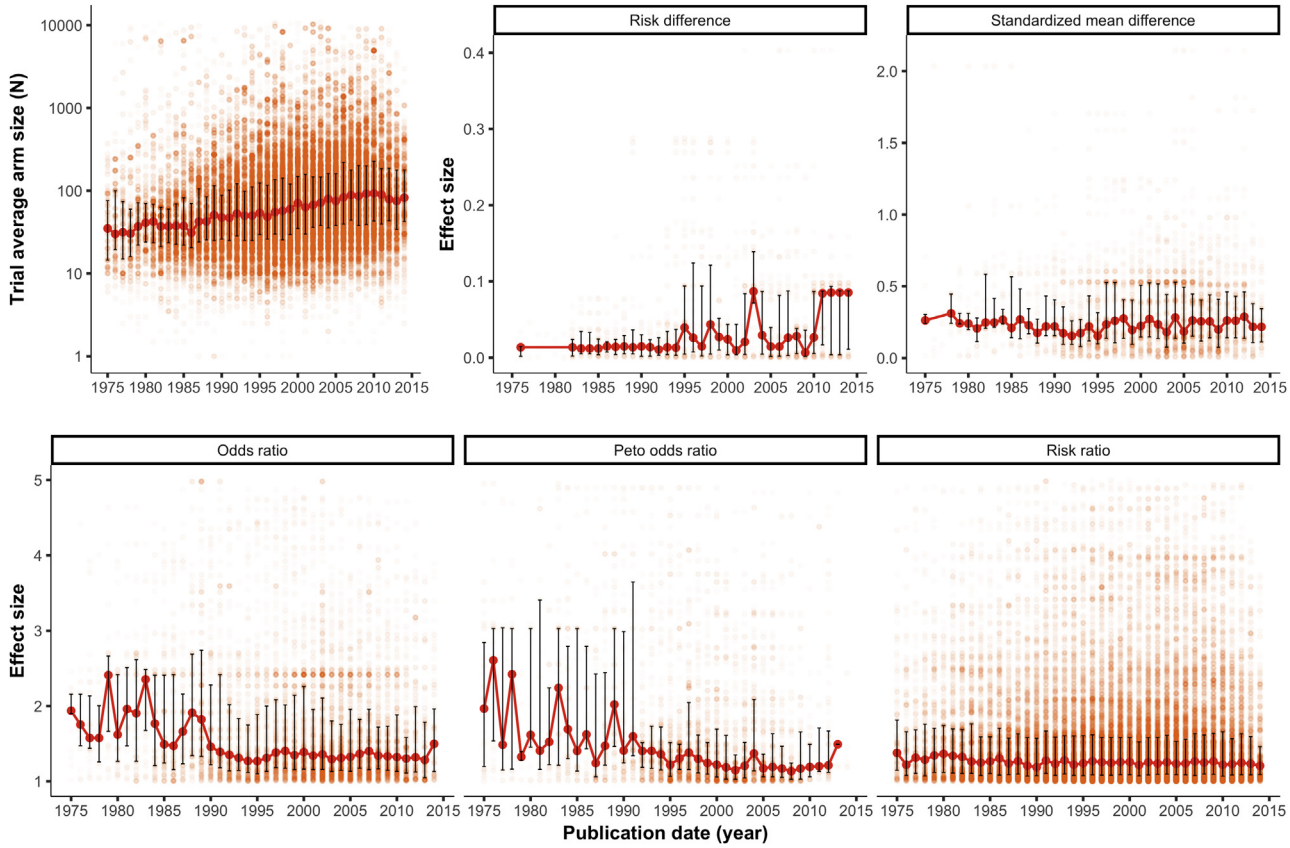

Fig. 2. The number of participants (*N*) enrolled in each trial arm, between 1975 and 2014, in red semitransparent dots (top left). Corresponding effect sizes—classified in Cochrane reviews as risk difference, standardized mean difference, (Peto) odds ratio or risk ratio—are shown in the remaining plots. Median and interquartile data are plotted annually. Years with less than ten studies with the specific measure were omitted from the plot. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

4. Discussion

The present study provides an overview of the statistical power in 136,212 clinical trials across all medical fields. Our analyses demonstrate that effect sizes are small, and that sample sizes of most clinical trials are too small to detect such an effect. Only 7% of individual trials had sufficient power to detect the observed effect from its respective meta-analysis. Although there is considerable room for improvement, an encouraging trend is the number of trials with sufficient power that has increased over 4 decades from 5% to 9%, and from 9% to 15% in trials from significant meta-analyses. On average, sample sizes have doubled between 1975 and 2014, whereas effect sizes did not increase over time.

Table 2. Median effect sizes for all meta-analyses

Reported effect measure	N meta-analyses (*n* included trials)	Raw effect size: median (interquartile range)	Standard effect size[a]
Odds ratio	1,798 (17,772)	1.37 (1.15–1.93)	0.09 (0.04–0.22)
Peto odds ratio	783 (8,421)	1.31 (1.11–1.79)	
Risk ratio	8,459 (100,534)	1.26 (1.09–1.64)	
Risk difference	187 (2,275)	0.02 (0.00–0.06)	
Standardized mean difference	625 (7,210)	0.20 (0.11–0.40)	0.20 (0.11–0.40)

Median effect sizes are computed based on the meta-analysis; every meta-analysis is taken into account once irrespective of the number of included trials. To obtain a meaningful summary statistic, effect sizes were transformed to be unidirectional: the absolute number of risk differences and standardized mean differences was taken, and for (Peto) odds ratios and risk ratio's effects below one were inversed (1 divided by the effect, e.g., an RR of 0.5 becomes 2.0). These transformations only change the direction and not the magnitude of the effect.

N = number of meta-analyses (number of included studies).

[a] Standard effect size: Cohen's d or h.

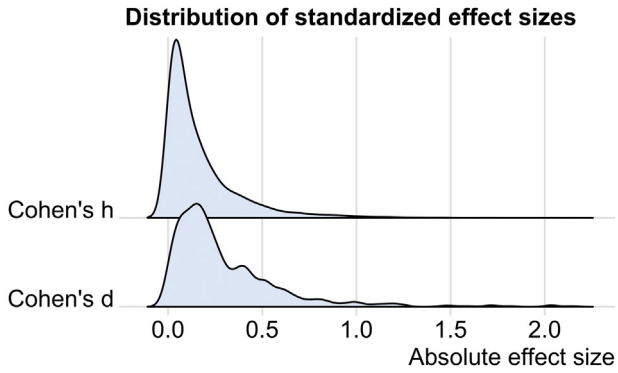

Fig. 3. Distribution plot of standardized effect sizes. Cohen's h was based on the proportion of events in the meta-analysis in case of dichotomous study outcomes. In studies comparing means the standardized mean difference (Cohen's d) of the meta-analysis was directly available in the Cochrane database.

The distribution of effect sizes (with a median Cohen's h of 0.09 and a median Cohen's d of 0.20) shows that large effects are rare. This information should be taken into account when designing a clinical trial and determining the required minimum sample size. The effect size summary statistics provided here could also be used as standard prior in Bayesian modeling in medical research because they are based on many thousands of trials covering the general medical field.

Our results are in agreement with a study by Turner et al, in which they also used the Cochrane database of systematic reviews (2008 version) to describe study power in clinical trials [3]. This study also showed low study power with a bimodal pattern of many low-powered studies and a small proportion of well-powered studies. The Turner study demonstrated a median power of 8%, whereas we find a comparable median power of 9% across all meta-analyses. This slightly higher percentage could be explained by the inclusion of more recent high-powered studies, or the exclusion of meta-analyses with less than five trials.

Our use of meta-analytic effect sizes to compute study power has two important shortcomings. First, although it is a fair—and the only available—approximation of the true effect of a given therapy, power and sample size calculations are designed to be performed a priori. We would fully endorse that for an individual study, there is no space for a post hoc power computation. Second, it may be questioned whether statistical power can be computed when the estimation of the effect size includes a null effect in the 95% CI. If there is no effect, a power calculation cannot be performed. If the null hypothesis "there is no effect" cannot be rejected, there is no clear effect size estimation available as the basis for the power calculation. We have therefore also included all results for the subset of significant meta-analyses.

By analyzing the temporal pattern across 4 decades, we identified an increase of study power over time. Moreover, because effect size estimates remained stable across time, our study clearly shows the need to increase sample sizes to design well-powered studies. A study on sample sizes determined in preregistration on ClinicalTrials.gov between 2007 and 2010 showed that over half of the registered studies included a required sample of 100 participants or less in their protocol [17]. We found that, within the published trials that have been included in a Cochrane meta-analysis, the findings are in line with these results, and although the average sample size has doubled since the 1970's, and median sample size in 2010–2014 was between 150 and 180.

An argument in defense of performing small (or underpowered) studies has been made based on the idea that small studies can be combined in a meta-analysis to increase power. Halpern et al already explained the invalidity of this argument in 2002 [1], most importantly because small studies are more likely to produce results with wide CIs and large *P*-values, and thus are more likely to remain unpublished. An additional risk of conducting uninformative studies is that a lack of an effect due to low power might decrease the interest by other research teams to examine the same effect. A third argument against performing small studies is given in a study by Nuijten et al [7], which indicates that the addition of a small, underpowered study to a meta-analysis may actually increase the bias of an effect size instead of decreasing it.

There are several limitations to consider in the interpretation of our results. First, the outcome parameter studied in the meta-analysis may be different than the primary outcome of the original study; it may have been adequately powered for a different outcome parameter. This could result in lower estimates of average power, although it seems unlikely that the average effect size of the primary outcomes is higher than the effect sizes in the Cochrane database. Second, by contrast, effect sizes from meta-analyses are considered to be an overestimation of the true effect because of publication bias [7,18]. Finally, in determining the required power for a study a "one size fits all" principle does not necessarily apply as Schulz and Grimes [19] also argue. However, although conventions are always arbitrary [12] a cutoff for sufficient power at 80% is reasonable.

With statistical power consistently increasing over time, our data offer perspective and show that we, the scientific community, are heading in the right direction. Nevertheless, it is clear that most clinical trials remain underpowered. Although there may be exceptions justifying small clinical trials, we believe that in most cases, underpowered studies are problematic. Clinical trials constitute the backbone of evidence-based medicine, and individual trials would ideally be interpretable in isolation, without waiting for a future meta-analysis. To further improve the current situation, trial preregistrations could include a mandatory section justifying the sample size, based on realistic expectations of the effect size, and preferably with explicit reference to earlier published results in the same field. If no prior literature exists for the specific condition or treatment,

we recommend that small effects should be assumed. Large-scale collaborations with the aim of performing either a multicenter study or a prospective meta-analysis may also increase sample sizes when individual teams lack the resources to collect larger sample sizes. Another important way to introduce long-lasting change is by improving the statistical education of current and future scientists [5]. Even though our analyses demonstrate that sufficient power in clinical trials is still problematic, the situation seems to be slowly improving. Together, these results encourage further efforts to increase statistical power in clinical trials to guarantee rigorous and reproducible evidence-based medicine.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2018.06.014.

References

[1] Halpern SD, Karlawish JHT, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA 2002;288:358−62.
[2] Rosoff PM. Can underpowered clinical trials be justified? IRB 2004; 26(3):16−9.
[3] Turner RM, Bird SM, Higgins JPT. The impact of study size on meta-analyses: examination of underpowered studies in cochrane reviews. PLoS One 2013;8:1−8.
[4] Szucs D, Ioannidis JPA. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLoS Biol 2017;15(3):1−18.
[5] Crutzen R, Peters GJY. Targeting next generations to change the common practice of underpowered research. Front Psychol 2017;8(1184):1−4.
[6] OpenScienceCollaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. Science 2015;349(6251): aac4716.
[7] Nuijten MB, Assen Van MALM, Veldkamp CLS, Wicherts JM. The replication paradox: combining studies can decrease accuracy of effect size estimates. Rev Gen Psychol 2015;19(2):172−82.
[8] Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 2013;14: 365−76.
[9] Ioannidis JPA, Stanley TD, Doucouliagos H. The power of bias in economics research. Econ J 2017;127(605):F236−65.
[10] Dumas-mallet E, Button KS, Boraud T, Gonon F, Munafò MR. Low statistical power in biomedical science: a review of three human research domains. R Soc Open Sci 2017;4:160254.
[11] Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Control Clin Trials 1981;2:93−113.
[12] Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Lawrence Earlbaum Associates; 1988.
[13] Otte WM. Temporal RCT power. Open science framework March 4 [Internet] 2017. Available at: https://osf.io/ud2jw/. Accessed June 29, 2018.
[14] Stanley TD, Doucouliagos H. Neither fixed nor random: weighted least squares meta-analysis. Stat Med 2015;34:2116−27.
[15] Stanley TD, Doucouliagos H, Ioannidis JPA. Finding the power to reduce publication bias. Stat Med 2017;36:1580−98.
[16] Champely S. pwr version 1.2_2 [Internet] 2018. Available at: http://cran.r-project.org/web/packages/pwr/.
[17] Califf RM, Zarin DA, Kramer JM, Sherman RE, Aberle LH, Tasneem A. Characteristics of clinical trials registered in ClinicalTrials.gov, 2007-2010. JAMA 2012;307:1838−47.
[18] Pereira TV, Ioannidis JPA. Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. J Clin Epidemiol 2011;64:1060−9.
[19] Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet 2005;365:1348−53.