

 Open access • Posted Content • DOI:10.1101/225730

Statistical power of clinical trials has increased whilst effect size remained stable: an empirical analysis of 137 032 clinical trials between 1975-2017 — Source link

Herm J. Lamberink, Willem M. Otte, Michel R.T. Sinke, Daniel Lakens ...+3 more authors

Institutions: [Utrecht University](#), [Eindhoven University of Technology](#), [Bond University](#), [VU University Amsterdam](#)

Published on: 28 Nov 2017 - [bioRxiv](#) (Cold Spring Harbor Laboratory)

Topics: [Sample size determination](#), [Randomized controlled trial](#), [Clinical trial](#), [Systematic review](#) and [Statistical power](#)

Related papers:

- [Statistical power of clinical trials increased while effect size remained stable : an empirical analysis of 136,212 clinical trials between 1975 and 2014](#)
- [Caution regarding the choice of standard deviations to guide sample size calculations in clinical trials](#)
- [Obtaining evidence by a single well-powered trial or several modestly powered trials:](#)
- [About half of the noninferiority trials tested superior treatments: a trial-register based study.](#)
- [Meta-analyses of cluster randomization trials. Power considerations.](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/statistical-power-of-clinical-trials-has-increased-whilst-4s9v0lh54s>

1 **Title: Statistical power of clinical trials has increased whilst effect size remained stable: an**
2 **empirical analysis of 137 032 clinical trials between 1975-2017**

3 *Short title:* Statistical power in clinical trials over time

4
5 **Authors**

6 Herm J Lamberink ^{a#}, Willem M Otte ^{a,b#*}, Michel RT Sinke ^b, Daniël Lakens ^c, Paul P Glasziou
7 ^d, Joeri K Tijdkink ^e, and Christiaan H Vinkers ^f

8 [#] Authors contributed equally

9 **Affiliations**

10 ^a Department of Child Neurology, Brain Center Rudolf Magnus, University Medical Center
11 Utrecht and Utrecht University, P.O. Box 85090, 3508 AB, Utrecht, The Netherlands

12 ^b Biomedical MR Imaging and Spectroscopy group, Center for Image Sciences, University
13 Medical Center Utrecht and Utrecht University, Heidelberglaan 100, 3584 CX Utrecht, The
14 Netherlands

15 ^c School of Innovation Sciences, Eindhoven University of Technology, Den Dolech 1, 5600 MB
16 Eindhoven, The Netherlands

17 ^d Centre for Research in Evidence-Based Practice, Faculty of Health Sciences and Medicine,
18 Bond University, Gold Coast, Queensland, Australia

19 ^e Department of Philosophy, VU University, De Boelelaan 1105, 1081 HV Amsterdam, The
20 Netherlands

21 ^f Department of Psychiatry, Brain Center Rudolf Magnus, University Medical Center Utrecht and
22 Utrecht University, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands

23

24 **Corresponding author**

25 Herm J Lamberink, Department of Child Neurology, Brain Center Rudolf Magnus, University
26 Medical Center Utrecht and Utrecht University, Room KC 03.063.0, P.O. Box 85090, 3508 AB,
27 Utrecht, The Netherlands. E: h.j.lamberink@umcutrecht.nl, T: +31 88 755 6030

28 **Abstract**

29 **Background.** Biomedical studies with low statistical power are a major concern in the scientific
30 community and are one of the underlying reasons for the reproducibility crisis in science. If
31 randomized clinical trials, which are considered the backbone of evidence-based medicine, also
32 suffer from low power, this could affect medical practice.

33 **Methods.** We analysed the statistical power in 137 032 clinical trials between 1975 and 2017
34 extracted from meta-analyses from the Cochrane database of systematic reviews. We determined
35 study power to detect standardized effect sizes according to Cohen, and in meta-analysis with p-
36 value below 0.05 we based power on the meta-analysed effect size. Average power, effect size
37 and temporal patterns were examined.

38 **Results.** The number of trials with power $\geq 80\%$ was low but increased over time: from 9% in
39 1975–1979 to 15% in 2010–2014. This increase was mainly due to increasing sample sizes,
40 whilst effect sizes remained stable with a median Cohen's d of 0.21 (IQR 0.12-0.36) and a
41 median Cohen's d of 0.31 (0.19-0.51). The proportion of trials with power of at least 80% to
42 detect a standardized effect size of 0.2 (small), 0.5 (moderate) and 0.8 (large) was 7%, 48% and
43 81%, respectively.

44 **Conclusions.** This study demonstrates that sufficient power in clinical trials is still problematic,
45 although the situation is slowly improving. Our data encourages further efforts to increase
46 statistical power in clinical trials to guarantee rigorous and reproducible evidence-based
47 medicine.

48 **Key words:** statistical power; clinical trial; randomized

49 **Key messages**

- 50 • Study power in clinical trials is low: 12% of trials are sufficiently powered (≥ 0.8) and
51 23% have a power above 0.5.
- 52 • Study power has increased from 9% in 1975–1979 to 15% in 2010–2014.
- 53 • Average effect sizes are low and did not increase over time.
- 54 • When determining the required sample size of a clinical trial, moderate effects should be
55 assumed to ensure an adequate sample size.

56 **Introduction**

57 The practice of conducting scientific studies with low statistical power has been consistently
58 criticized across academic disciplines (1–5). Statistical power is the probability that a study will
59 detect an effect when there is a true effect to be detected. Underpowered studies have a low
60 chance of detecting true effects and have been related to systematic biases including inflated
61 effect sizes and low reproducibility (6, 7). Low statistical power has been demonstrated, amongst
62 others, in the fields of neuroscience and psychology (4, 8, 9). For clinical trials in the field of
63 medicine, the issue of sample size evaluation and statistical power is essential since clinical
64 decision making and future research are based on these clinical trials (10, 11). Moreover, low
65 power in clinical trials may be unethical in light of the low informational value from the outset
66 while exposing participants to interventions with possible negative (side) effects (1). Also in
67 medical research statistical power is low (3, 8), but a systematic overview of temporal patterns of
68 power, sample sizes, and effect sizes across medical fields does not exist. In the current study,
69 we provide a comprehensive overview of study power, sample size, and effect size estimates of
70 clinical trials published since 1975 which are included in the Cochrane database of systematic
71 reviews, and analyse emerging trends over time.

72 **Materials and Methods**

73 Data were extracted and calculated from trials included in published reviews from the second
74 Issue of the 2017 Cochrane Database of Systematic Reviews. Cochrane reviews only include
75 meta-analyses if the methodology and outcomes of the included trials are comparable in cross
76 study populations. Meta-analysis data is available for download in standardized XML-format for
77 those with an institutional Cochrane Library license. We provide open-source software to
78 convert these data and reproduce our entire processing pipeline (15).

79 Trials were selected if they were published after 1974 and if they were included in a meta-
80 analysis based on at least five trials. For each individual clinical trial, publication year, outcome
81 estimates (odds or risk ratio, risk difference or standardized mean difference) and group sizes
82 were extracted. The power of individual trials was first calculated for detecting small, medium
83 and large effect sizes (Cohen's d or h of 0.2, 0.5 and 0.8, respectively); (11), based on the sample
84 sizes in both trial arms, using a 5% α threshold. Next, analyses were performed based on the
85 observed effect size from meta-analyses with a p-value below 0.05, irrespective of the p-value of
86 the individual trial; if a meta-analysis has a p-value higher than 0.05, the null-hypothesis "there
87 is no effect" cannot be discarded, and power cannot be computed for absent effects. Given that
88 publication bias inflates meta-analytic effect size estimates (7, 13), this can be considered a
89 conservative approach. All analyses were carried out in R using the 'pwr' package (16).

90 Following minimum recommendations for the statistical power of studies (11), comparisons with
91 a power above or equal to 80% were considered to be sufficiently powered. Study power, group
92 sizes and effect sizes over time were summarised and visualized for all clinical trials.

93 **Results**

94 Data from 137 032 clinical trials were available, from 11 852 meta-analyses in 1918 Cochrane
95 reviews. Of these trials 8.1% had a statistical power of at least 80% (the recommended minimum
96 (11), which we shall denote as ‘sufficient power’) to detect a small effect (Cohen’s d or h 0.2),
97 48% and 81% of trials had sufficient power to detect a moderate (Cohen’s d or h 0.5) or large
98 effect (Cohen’s d or h 0.8), respectively (Figure 1). This figure shows that there was no
99 difference between trials included in meta-analyses with a p -value below 0.05 and those above
100 this threshold.

101 <Figure 1 here>

102 To compute study power to detect the effect observed in the meta-analysis, we examined the
103 subset of meta-analyses with overall group differences with a p -value <0.05 : 78 281 trials
104 (57.1%) from 5903 meta-analyses (49.8%) in 1411 Cochrane reviews (73.6%). All following
105 results are based on this sub-selection of meta-analyses. On average, 12.5% of these trials were
106 sufficiently powered to detect the observed effect size from the respective meta-analysis (Figure
107 1). The median (interquartile range, IQR) power for the four categories corresponding to Figure
108 1 was 19% (12-37%), 78% (49-98%), 99% (87-100%) and 20% (10-48%), respectively.

109 Between 1975-1979 and 2010-2014 study power increased, with the median rising from 16%
110 (IQR 10-39) to 23% (IQR 12-55) (Figure 2, left), and the proportion of sufficiently powered
111 studies rose from 9.0% (95% confidence interval (CI) for proportions 7.6-10.6) to 14.7% (95%
112 CI 13.9 - 15.5) (Figure 2, top right). This trend is also seen across medical disciplines
113 (Supplementary Figure 2). When the power threshold is set at a minimum of 50% power, the
114 proportion of trials with sufficient power is still low but also rising: from 19.3% (95% CI 17.3-

115 21.4) in the late 1975-1979 to 27.5% (95% CI 23.5-31.9) in 2010-2014 (Supplementary Figure
116 1). The distribution of power showed a bimodal pattern, with many low-powered studies and a
117 small peak of studies with power approaching 100% (Figure 2, bottom right).

118 <Figure 2 here>

119 The average number of participants enrolled in a trial arm increased over time (Figure 3, top
120 left). The median group size in 1975-1979 ranged between 33 and 45; for the years 2010-2014
121 the median group size was between 74 and 92. The median effect sizes are summarized in Table
122 1; these remained stable over time (Figure 3). The standardized effect sizes were small to
123 moderate, with a median Cohen's h of 0.21 (IQR 0.12-0.36) and a median Cohen's d of 0.31
124 (0.19-0.51) (Table 1); Supplementary Figure 3 shows the distribution plots for these two
125 measures.

126 <Figure 3 here>

127 <Table 1 here>

128 **Discussion**

129 Our study provides an overview of the statistical power in 137 032 clinical trials across all
130 medical fields. This analysis suggests most clinical trials are too small: the majority had
131 insufficient power to detect small or moderate effect sizes, whereas most observed effects in the
132 meta-analyses were actually small to moderate. Only 12% of individual trials had sufficient
133 power to detect the observed effect from its respective meta-analysis. This study adds to the
134 existing evidence that low statistical power is a widespread issue across clinical trials in
135 medicine (3). Though there is considerable room for improvement, an encouraging trend is the
136 number of trials with sufficient power has increased over four decades from 9% to 15%. On
137 average, sample sizes have doubled between 1975 and 2017 whereas effect sizes did not increase
138 over time.

139 The average effect sizes were small, with a median Cohen's d of 0.21 and a median Cohen's h of
140 0.31. These results show that large effects are rare, which should be taken into account when
141 designing a study and determining the required minimum sample size. The effect size summary
142 statistics provided here could also be used as standard prior in Bayesian modelling in medical
143 research, since they are based on many thousands of trials covering the general medical field.
144 The study by Turner et al. (3) also used the Cochrane library (the 2008 edition) to investigate
145 power. They showed low power of clinical trials, and a bimodal distribution of statistical power
146 with many low-powered studies and a small peak of high-powered studies; a result also shown in
147 neurosciences (8) and replicated by our analysis. By analysing the temporal pattern across four
148 decades, we have been able to identify an increase of study power over time. Moreover, since
149 effect size estimates remained stable across time, our study clearly shows the need to increase
150 sample sizes to design well-powered studies. A study on sample sizes determined in

151 preregistration on ClinicalTrials.gov between 2007-2010 showed that over half of the registered
152 studies included a required sample of 100 participants or less in their protocol (12). Our findings
153 are in line with these results, and although the average sample size has doubled since the 1970's,
154 we found that the median sample size in the 2010's was still below 100.

155 An argument in defence of performing small (or underpowered) studies has been made based on
156 the idea that small studies can be combined in a meta-analysis to increase power. Halpern and
157 colleagues already explained why this argument is invalid in 2002 (1), most importantly because
158 small studies are more likely to produce results with wide confidence intervals and large p-
159 values, and thus are more likely to remain unpublished. An additional risk of conducting
160 uninformative studies is that a lack of an effect due to low power might decrease the interest by
161 other research teams to examine the same effect. A third argument against performing small
162 studies is given in a study by Nuijten and colleagues (7), which indicates that the addition of a
163 small, underpowered study to a meta-analysis may actually increase the bias of an effect size
164 instead of decreasing it.

165 There are several limitations to consider in the interpretation of our results. First, the outcome
166 parameter studied in the meta-analysis may be different than the primary outcome of the original
167 study; it may have been adequately powered for a different outcome parameter. This could result
168 in lower estimates of average power, although it seems unlikely that the average effect size of the
169 primary outcomes is higher than the effect sizes in the Cochrane database. Second, in contrast,
170 effect sizes from meta-analyses are considered to be an overestimation of the true effect because
171 of publication bias (7, 13). Lastly, in determining the required power for a study a 'one size fits
172 all' principle does not necessarily apply as Schulz & Grimes (14) also argue. However, although
173 conventions are always arbitrary (11) a cut-off for sufficient power at 80% is reasonable.

174 With statistical power consistently increasing over time, our data offer perspective and show that
175 we are heading in the right direction. Nevertheless, it is clear that many clinical trials remain
176 underpowered. Although there may be exceptions justifying small clinical trials, we believe that
177 in most cases underpowered studies are problematic. Clinical trials constitute the backbone of
178 evidence-based medicine, and individual trials would ideally be interpretable in isolation,
179 without waiting for a future meta-analysis. To further improve the current situation, trial pre-
180 registrations could include a mandatory section justifying the sample size, either based on a
181 sufficient power for a smallest effect size of interest, or the precision of the effect size estimate.
182 Large-scale collaborations with the aim of performing a either a multi-centre study or a
183 prospective meta-analysis might also help to increase sample sizes when individual teams lack
184 the resources to collect larger sample sizes. Another important way to introduce long-lasting
185 change is by improving the statistical education of current and future scientists (5).

186

187 **Funding**

188 This work was supported by The Netherlands Organisation for Health Research and Development
189 (ZonMW) grant “Fostering Responsible Research Practices” (445001002).

190

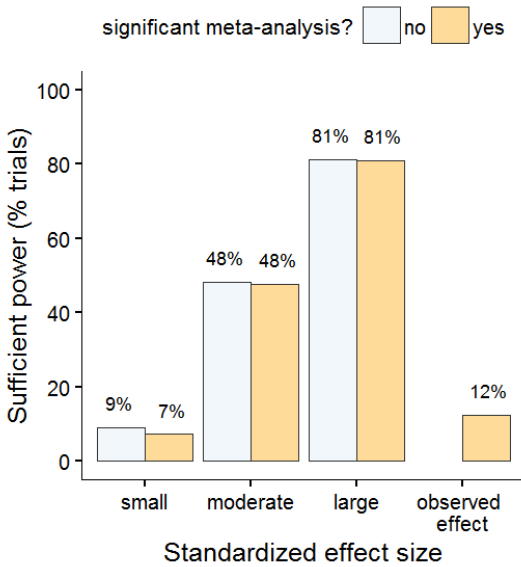
191 **References**

- 192 1. Halpern SD, Karlawish JHT, Berlin JA (2002) The continuing unethical conduct of
193 underpowered clinical trials. *JAMA* 288(3):358–362.
- 194 2. Rosoff PM (2004) Can Underpowered Clinical Trials Be Justified? *IRB Ethics Hum Res*
195 26(3):16–19.
- 196 3. Turner RM, Bird SM, Higgins JPT (2013) The Impact of Study Size on Meta-analyses :
197 Examination of Underpowered Studies in Cochrane Reviews. *PLoS One* 8(3):1–8.
- 198 4. Szucs D, Ioannidis JPA (2017) Empirical assessment of published effect sizes and power
199 in the recent cognitive neuroscience and psychology literature. *PLoS Biol* 15(3):1–18.
- 200 5. Crutzen R, Peters GJY (2017) Targeting next generations to change the common practice
201 of underpowered research. *Front Psychol* 8(1184):1–4.
- 202 6. Open Science Collaboration (2015) Estimating the reproducibility of psychological
203 science. *Science (80-)* 349(6251):aac4716.
- 204 7. Nuijten MB, Assen MALM Van, Veldkamp CLS, Wicherts JM (2015) The Replication
205 Paradox : Combining Studies can Decrease Accuracy of Effect Size Estimates. *Rev Gen*
206 *Psychol* 19(2):172–182.
- 207 8. Button KS, Ioannidis JPA, Mokrysz C, et al. (2013) Power failure: why small sample size
208 undermines the reliability of neuroscience. *Nat Rev Neurosci* 14:365–376.
- 209 9. Dumas-mallet E, Button KS, Boraud T, Gonon F, Munafò MR (2017) Low statistical
210 power in biomedical science : a review of three human research domains. *R Soc open sci*
211 4:160254.
- 212 10. Lachin JM (1981) Introduction to sample size determination and power analysis for
213 clinical trials. *Control Clin Trials* 2:93–113.

- 214 11. Cohen J (1988) *Statistical power analysis for the behavioral sciences* (Lawrence
215 Earlbaum Associates, Hillsdale, NJ). 2nd ed.
- 216 12. Califf RM, Zarin DA, Kramer JM, et al. (2012) Characteristics of Clinical Trials
217 Registered in ClinicalTrials.gov, 2007-2010. *JAMA* 307(17):1838–1847.
- 218 13. Pereira T V, Ioannidis JPA (2011) Statistically significant meta-analyses of clinical trials
219 have modest credibility and inflated effects. *J Clin Epidemiol* 64(10):1060–1069.
- 220 14. Schulz KF, Grimes DA (2005) Sample size calculations in randomised trials : mandatory
221 and mystical. *Lancet* 365:1348–1353.
- 222 15. Otte WM (2017) Temporal RCT power. Open Science Framework March 4. Available at:
223 <https://osf.io/ud2jw/>.
- 224 16. Champely S (2017) pwr. Available at: <http://cran.r-project.org/web/packages/pwr/>.
225

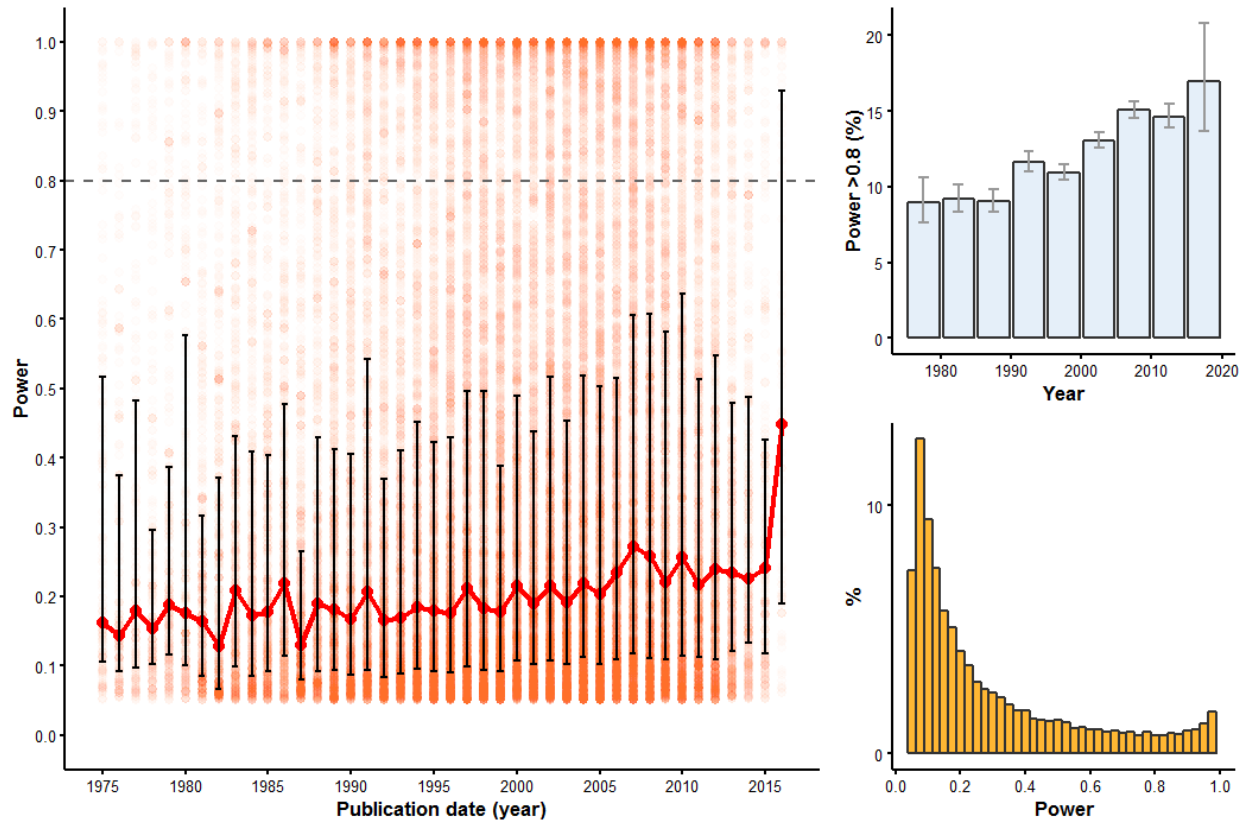
226 **Figure legends**

227



228

229 **Figure 1** | The proportion of trials that are sufficiently powered ($\geq 80\%$ power) for finding a
230 hypothetical effect equal to Cohen's d or h of 0.2 (small effect), 0.5 (moderate effect) and
231 0.8 (large effect) comparing trials from 'significant' (p-value < 0.05 , $n = 78,401$) and
232 'non-significant' ($n = 58,631$) meta-analyses, and the proportion of trials sufficiently
233 powered for finding the effect as observed in the respective significant meta-analysis.



234

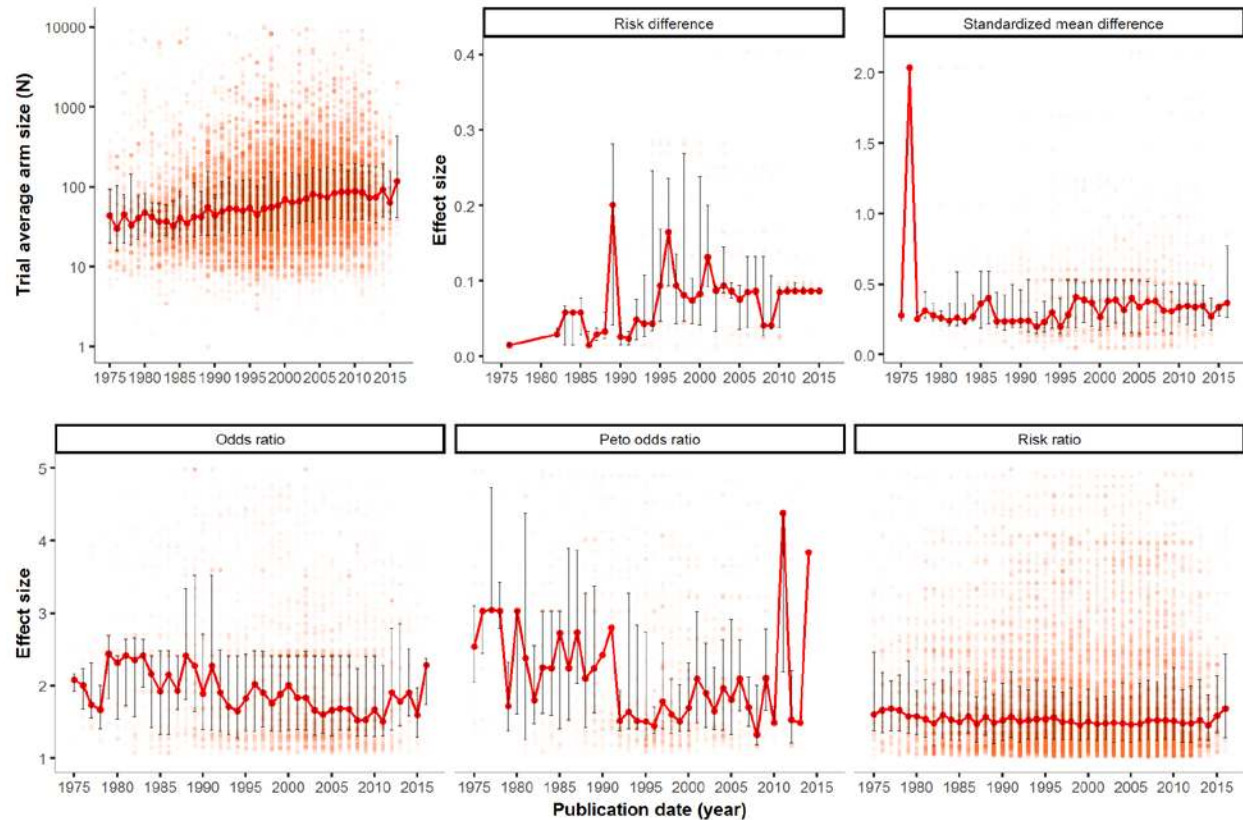
235 **Figure 2** | Statistical power of clinical trials between 1975 and 2017 from meta-analyses with p-
236 value <0.05 (left). Individual comparisons are shown as semi-transparent dots. Median
237 power is shown in red with interquartile range as error bars. The percentage of adequately
238 powered trial comparisons (i.e. $\geq 80\%$ power) is increasing over time (top right). The
239 biphasic power distribution of the trials in general is apparent (bottom right).

240

241

242

243



244

245 **Figure 3** | The number of participants (N) enrolled in each trial arm, between 1975 and 2017 in
246 red semi-transparent dots (top left). Corresponding effect sizes – classified in Cochrane
247 reviews as risk difference, standardized mean difference, (Peto) odds ratio or risk ratio –
248 are shown in the remaining plots. Median and interquartile data are plotted annually.

249

250

Table 1 | Median effect sizes for all meta-analyses with p-value <0.05

Reported effect measure	N meta-analyses (n included trials)	Median effect size (IQR)	Standard effect size*
Odds ratio	879 (10303)	1.83 (1.41-2.63)	} 0.21 (0.12-0.36)
Peto odds ratio	348 (3956)	1.82 (1.43-2.92)	
Risk ratio	4230 (57976)	1.57 (1.30-2.22)	
Risk difference	68 (996)	0.08 (0.04-0.14)	-
Standardized mean difference	378 (5170)	0.31 (0.19-0.51)	0.31 (0.19-0.51)

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)

*standard effect size: Cohen's d or h

-no standard effect size could be computed due to missing confidence intervals

251