BMJ Open

Assessing treatment effects and publication bias across different specialties in medicine: a meta-epidemiological study

Simon Schwab ,[1] Giuachin Kreiliger,[2] Leonhard Held[1]

To cite: Schwab S, Kreiliger G, Held L. Assessing treatment effects and publication bias across different specialties in medicine: a meta-epidemiological study. *BMJ Open* 2021;**11**:e045942. doi:10.1136/bmjopen-2020-045942

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-045942).

Received 21 October 2020
Accepted 09 August 2021

[1]Center for Reproducible Science & Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland
[2]Master Program in Biostatistics, University of Zurich, Zurich, Switzerland

Correspondence to
Dr Simon Schwab;
simon.schwab@uzh.ch

ABSTRACT

Objectives To assess the prevalence of statistically significant treatment effects, adverse events and small-study effects (when small studies report more extreme results than large studies) and publication bias (over-reporting of statistically significant results) across medical specialties.

Design Large meta-epidemiological study of treatment effects from the Cochrane Database of Systematic Reviews.

Methods We investigated outcomes from 57 162 studies from 1922 to 2019, and overall 98 966 meta-analyses and 5534 large meta-analyses (≥10 studies). Egger's and Harbord's tests to detect small-study effects, limit meta-analysis and Copas selection models to bias-adjust effect estimates and generalised linear mixed models were used to analyse one of the largest collections of evidence in medicine.

Results Medical specialties showed differences in the prevalence of statistically significant results of efficacy and safety outcomes. Treatment effects from primary studies published in high ranking journals were more likely to be statistically significant (OR=1.52; 95% CI 1.32 to 1.75) while randomised controlled trials were less likely to report a statistically significant effect (OR=0.90; 95% CI 0.86 to 0.94). Altogether 19% (95% CI 18% to 20%) of the large meta-analyses showed evidence for small-study effects, but only 3.9% (95% CI 3.4% to 4.4%) showed evidence for publication bias after further assessment of funnel plots. Adjusting treatment effects resulted in overall less evidence for efficacy.

Conclusions These results suggest that reporting of large treatment effects from small studies may cause greater concern than publication bias. Incentives should be created so that studies of the highest quality become more visible than studies that report more extreme results.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This is one of the few studies providing a large-scale assessment of treatment effects, small-study effects and publication bias across different medical specialties.
⇒ Many methods are available to assess small-study effects and publication bias. We only considered Egger's test and Harbord's test; however, these were among the most widely recommended ones.
⇒ Bias may be underestimated as the methods used in this study have low statistical power; thus, only a subset of large meta-analyses (≥10 studies) were considered.
⇒ The assessment of small-study effects is restricted to the traditional p<0.05 threshold. However, we additionally bias adjusted the effect estimates without any thresholding.
⇒ The restriction to large meta-analyses (≥10 studies) with low heterogeneity (I^2 <50) leads to selection bias, limits generalisability and is a limitation of this study.

INTRODUCTION

Publication bias is a major concern in clinical research as it affects the combined effects from meta-analyses of intervention studies and distorts the overall evidence for the efficacy of a treatment.[1 2] The problem was already characterised more than half a century ago[3] and a number of studies provided compelling evidence for publication bias by following studies from protocol approval until the publication of outcomes.[4 5] The problem originates when studies with a null result are not considered worthy to be written up and submitted by researchers, or may not be treated favourably in peer review; hence, are less likely to be published.[6 7] Moreover, statistically significant studies are more likely to be published in journals with a high citation impact factor[8] and get more citations.[9] Thus, statistically significant results receive more attention than null (non-significant) results.

Many countries and medical journal publishers require trial registration with the aim to prevent publication bias, for example, the International Committee of Medical Journal Editors. Since 2007 the US Food and Drug Administration Amendments Act (FDAAA) demands that interventional clinical trials report their results directly to the US trial registry 'ClinicalTrials.gov' within

1 year of the completion of the study; however, compliance with FDAAA 2007 is poor with 59% of the studies not reporting results in time.[10] The retrieval process of unpublished results can be difficult, time consuming and not always successful. Thus, statistical methods can be useful to not only assess but also adjust meta-analyses for inflated treatment effects and publication bias.

In clinical research multiple studies on the same treatment are conducted, and meta-analyses are the established statistical tool to estimate the combined treatment effect. Combined effects are essentially weighted averages of the study effects with effects from large studies having more weight than effects from smaller studies. If there is publication bias, smaller studies with null findings will be published less likely. This leads to a positive association between the SE of effect sizes and the effect size itself, known as the small-study effect.[11] This association is assessed by Egger's regression and can be shown as asymmetry in a funnel plot,[12] but this method also has some limitations.[13]

The estimate of the correlation between study size and effect size may be erroneous if there are few studies, or if there is heterogeneity across studies. Even though it is often assumed that publication bias may be a plausible explanation for small-study effects, there are also other possible causes. For example, low-quality studies that report inflated effect sizes, or clinical heterogeneity of patients when small studies focus on high-risk patients for whom the treatment may be more effective. Therefore, it is important to follow guidelines when applying such methods[14 15] such as excluding meta-analyses with high heterogeneity or less than 10 studies as they can mislead small study tests. Sterne *et al*[14] recommended Harbord's test[16] for dichotomous outcomes using ORs and Egger's test[12] for continuous outcomes. Also, the Cochrane handbook recommends the same tests, provided that there are at least 10 studies.[17]

In contrast to regression methods that assess small-study effects, selection models explicitly model publication bias. Selection models adjust meta-analytical data by specifying a model that describes the mechanism by which effect sizes may be suppressed. The Copas selection model[18] is among the more sophisticated selection models[19] and investigates whether studies with a certain effect size have a greater probability to enter a meta-analysis. A common feature in both regression-based methods and the Copas selection model[20] is that effect estimates can be adjusted for bias. For example, regression-based adjustments of treatment effects were successful in the prediction of the effect of antidepressant trials from the FDA trial registry using a biased subset of the data.[21]

There have been other studies on the extent of publication bias in the Cochrane Database of Systematic Reviews (CDSR). Sutton *et al*[22] estimated that around 50% of 48 Cochrane reviews had missing studies. A large study[15] investigated meta-analyses with binary outcomes and found small-study effects in 19% of the 366 meta-analyses. Contrarily, another study found no convincing evidence for publication bias[23] but they only investigated a small subset of 83 meta-analyses. A study that used a selection model with 1106 meta-analyses found positive findings to be 27% more likely to be included in a meta-analysis.[24] An extensive and recent study by Lin *et al*[25] investigated 30 000 meta-analyses from the CDSR and tested various methods among those regression-based tests (but no selection models) and reported considerable small-study effects in 20%–40% of the meta-analyses; however, they included meta-analysis with five studies which is at odds with the common guideline of at least 10 studies. Some methodological studies have been performed with data from the CDSR to evaluate various asymmetry tests and found that Egger's linear regression was most sensitive to detect asymmetry.[26 27] Statistical power can be increased by choosing a more liberal p value threshold with no substantial increase in false positive rate,[28] and indeed in many studies a significance level of $p<0.10$ was used.

Previous studies were sometimes constrained to only a subset of the CDSR data or to methodology that is generally not recommended.[29] Often it was not reported how outcomes related to efficacy were distinguished from outcomes related to adverse events. This is important because publication bias may not operate in favour of statistically significant adverse effects. Furthermore, adjustment of the effect sizes is rarely performed, and to the best of our knowledge, none of the studies compared small-study effects and publication bias across different medical specialties.

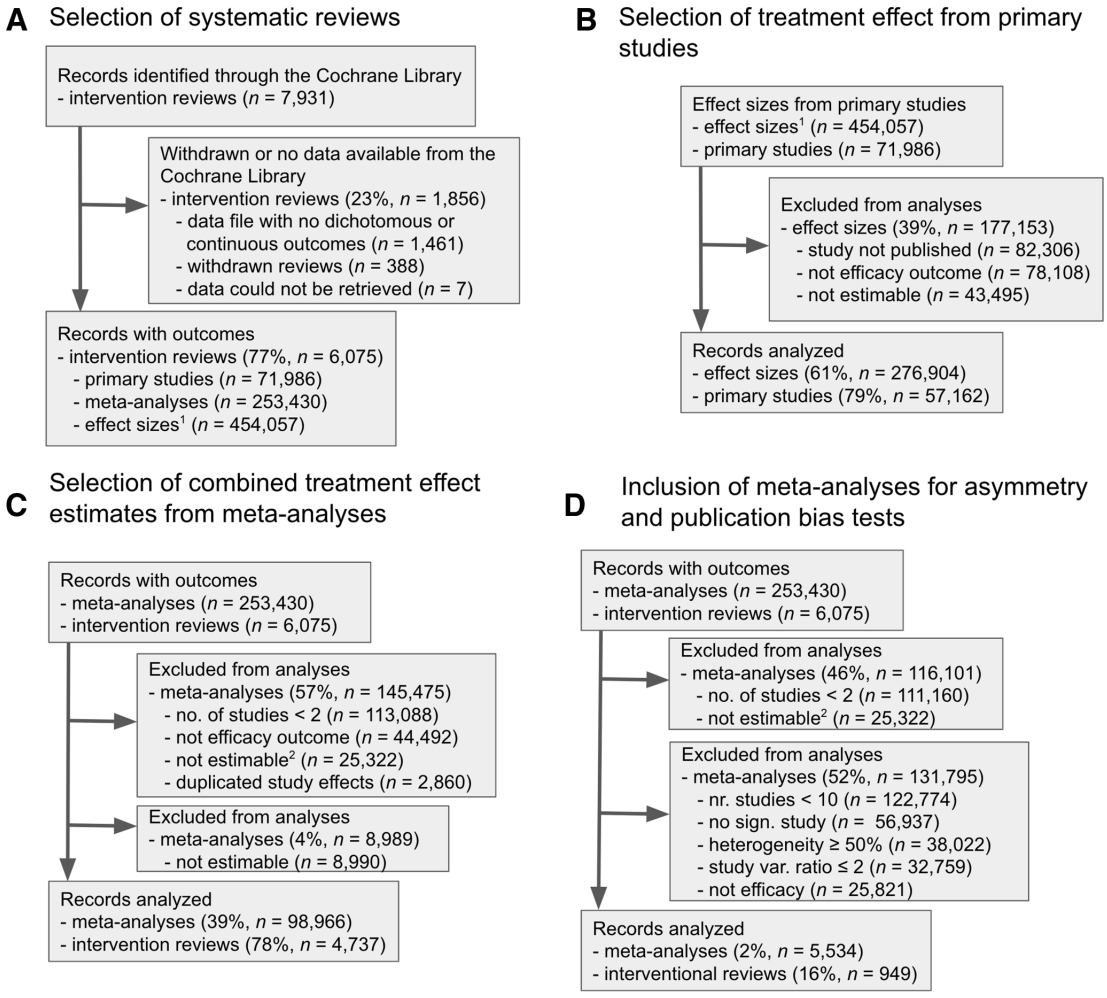
The aim of this study was a large-scale meta-epidemiological assessment of the prevalence of statistically significant treatment effects, adverse events, small-study effects and publication bias in meta-analyses across different medical specialties. The four objectives are summarised in more details as follows:

1. Assessment of the reported statistically significant effect sizes in primary studies.
2. Assessment of the statistically significant combined effect estimates in meta-analyses.
3. Estimation of small-study effects and publication bias in large meta-analyses.
4. Statistical bias adjustment of treatment effect estimates.

The hypotheses were specified in more detail in the protocol (osf.io/3a28k) and in the online supplemental table 1.

METHODS

A study protocol was written, and the methods were specified in advance including a statistical analysis plan. The protocol was registered on 9 July 2019 (osf.io/3a28k). Deviations from the protocol are specified at the end of the methods sections. The complete pipeline with analysis code written in R is available at osf.io/uv397. We reported this study following the guidelines for meta-epidemiological studies[30] as well as the Strengthening the Reporting of ObservationalStudies in Epidemiology



[1]Effect sizes unique in a medical specialty; the effect sizes unique across specialties were *n* = 453,397
[2]Not estimable as specified in the CDSR

Figure 1 Selection process from the Cochrane Database of Systematic Reviews (CDSR). (A) Data from 77% of the systematic reviews were retrieved. Selection of treatment effects from primary studies (B) and selection of meta-analyses (C–D).

(STROBE) statement,[31] the STROBE checklist for this study can be found at osf.io/b6gxf.

Search strategy and study selection

Systematic reviews of the type 'intervention' were selected from the CDSR which were 7931 reviews as of 11 November 2019. All reviews involved an intervention, for example, a drug, a surgical procedure, a psychotherapy, a medical device, preventive care, etc.

Data collection

Data from 6075 (77%) of all the systematic reviews could be retrieved which comprised 253 430 meta-analyses with outcomes from 71 986 primary studies from over 50 million patients (figure 1A). For each systematic review, the data were downloaded from the Cochrane Library as an XML file, parsed and aggregated in a large database in R. The database included the sample sizes, mean and SD (for continuous outcomes) or the number of events (for dichotomous outcomes) for both the intervention and control arm. We further extended the dataset by scraping additional information for the 6075 Cochrane reviews: the full reference of the included primary studies (including the journal name) and the table 'characteristics of included studies' which included additional details on the methods, participants, interventions and outcomes of the intervention studies. We collected citation information from Google Scholar (May 2020; scholar.google.com) for 40 306 (70%) and the Scimago journal ranks (SJR; February 2020; www.scimagojr.com) for 39 270 (69%) of the primary studies selected for analyses in figure 1B.

Data processing

Beside the outcomes related to the efficacy of an intervention, there were also outcomes related to adverse events (13.5%), withdrawal/dropout from the study (2.4%) and bias/sensitivity analyses (1.3%). Outcomes related to adverse events, withdrawal and bias were identified by using a set of keywords and regular expressions on the comparison name, outcome name and subgroup name as provided by the CDSR.

The 53 Cochrane review groups (which also included past groups) were merged into 19 medical specialties, see online supplemental table 2. Small groups were merged when they contained less than 150 reviews and assignment was based on thematic overlap, for example, various cancer groups into 'oncology', various neurological conditions into 'neurology'.

Inclusion criteria

The data selection process for the analyses conducted is shown in figure 1B–D. The analysis of the effect sizes (figure 1B) included all effects from published primary studies. For the analysis of the combined effects from meta-analyses the number of primary studies was at least 2 (figure 1C). We only considered study effects from published sources in the meta-analyses. Meta-analyses based on subgroups were generally included as they conveyed important evidence (eg, different comparison of drugs).

For the small-study effects tests, we followed the inclusion criteria by Ioannidis and Trikalinos.[15] Among them are at least 10 studies and low to moderate heterogeneity (I^2 <50%) in the meta-analysis (see figure 1D and protocol for details).

Statistical analyses

We used Wilson CIs for proportions and Wald CIs for ORs and regression coefficients. Differences in proportions across medical specialties were assessed with a Pearson's χ^2 test without continuity correction.

Effect sizes from primary studies

Effect sizes from primary studies included various outcome measures. The most common were risk ratios (46%), mean differences (26%), ORs (11%) and standardised mean differences (SMD) (8%). Comparing reported effect sizes from primary studies across medical specialties required harmonisation. We recalculated all effect sizes for the primary studies using escalc() from the R package metafor.[32] For continuous outcomes the SMD (Hedges' g), for dichotomous outcomes the ORs were used. ORs were transformed into Hedges' g. All effects on the common SMD scale were then transformed into Pearson's r.[33] An overview of the harmonisation of effect estimates from primary studies and meta-analyses and associated statistical analyses are shown in online supplemental figure 1. Statistical significance was assessed with a Wald test which was performed on the original effect measure as reported in the CDSR (eg, risk ratio, OR); for mean differences we applied a two-sample Student's t-test.

Combined effect estimates from meta-analyses

We used the R package meta[34] for conducting random effects meta-analyses (with DerSimonian-Laird estimator for between study variance) for continuous outcomes (Hedges' g) and binary outcomes (OR). The combined effect from the random effects meta-analysis was converted into Pearson's r[33] for comparison across specialties.

Small-study effect tests and adjustment of combined effects

Given the vast number of methods that exist to study publication bias and the lack of proper validation,[35] we restricted our analyses to Egger's[12] and Harbord's test.[16] For continuous outcomes we used Egger's regression of the effect estimates on their precision (ie, inverse SE). Harbor's test is a modification for ORs that resolves the issue of the standard errors not being independent from the effect estimates even in the absence of small-study effects. Meta-analyses of Peto ORs were performed with ORs, and meta-analysis of HRs or risk differences were performed with risk ratios instead; tests for risk differences are not recommended.[17] All tests were performed one sided (p<0.05); the rationale for this will be explained further below.

The metasens package[36] was used to adjust combined effects via a regression-based method known as limit meta-analysis[37 38] and the Copas selection model.[18 20] Both methods are based on a random effects model. The Copas selection model is an alternative to regression-based tests that explicitly models the chance of publication, that is, whether studies with a certain effect size have a greater probability to enter a meta-analysis. All adjustments were performed on the log ORs for dichotomous outcomes and on SMDs for continuous outcomes and were afterwards transformed to Pearson's r scale[33] in order to compare results across medical specialties.

We assessed whether the adjustment increased or decreased the evidence for an effect of the intervention. As we were not able to infer the direction of the anticipated treatment effect for every meta-analysis, we defined the direction in favour of the treatment as the sign with more statistically significant study results.[23] Once the anticipated sign of the treatment effect was determined, a one-sided regression-based test (p<0.05) could be performed which had the advantage to exclude situations where asymmetry was not caused by a small-study effect.

Small-study effect with probable publication bias

Small-study effect tests do not directly assess publication bias, and it is important to further evaluate the meta-analyses.[13] We chose the meta-analyses with significant small-study effects (at p<0.05) and determined the proportion of studies with a traditionally significant result vs a null result in any of the top 50%, 33% and 25% of the studies with the largest SEs (ie, among the smallest studies). If an equal or larger proportion of statistically significant studies were found in any of the quantiles, we assumed 'small-study effects with probable publication bias' as publication bias seemed more likely in such situations. While publication bias may theoretically still be possible and there are 'hidden' studies in the funnel plot, their impact (magnitude of bias) may not be so severe when a majority of the smaller studies that were published show a null result.

Change in evidence for treatment effects

We converted the p values from the combined treatment effects from a random effects meta-analysis, from the bias-adjusted regression-based analysis and from the Copas selection model analysis into Bayes factor bounds[39] and categorised those into six levels of 'weak' to 'decisive' evidence against the null hypothesis.[40] The change in evidence after adjustment was determined by comparing the proportions of treatment effects falling in those categories.

Modelling of outcomes: statistical significance, small-study effects and bias adjustment

Six generalised linear mixed-effects models (labelled 1–6 in the following text) were performed with the R package lme4[41] for the binary outcome statistical significance ($p<0.05$, two sided) of the treatment effects in primary studies (1) and combined effects from meta-analyses (2). We modelled the evidence for asymmetry using the test statistic (t-score) from the regression-based test (3) as well as the binary outcome (4) of whether the asymmetry tests gave a statistically significant result or not ($p<0.05$, one sided). We also modelled the change in Pearson's r after adjustment with regression-based adjustment (5) and Copas adjustment (6). The explanatory variables for models 1–6 were specified in the protocol.

Explanatory variables that were right-skewed were $\log_2$ transformed. Random effects accounted for dependencies and nesting in the data; these were the study identifier to model results from primary studies, and the Cochrane review identifier for the meta-analyses. The models were in more detail specified in the study protocol section 4.3.

Missing data

There were no missing data for the outcome variables: treatment effect from primary studies, combined treatment effect from meta-analyses, assessment of small-study effect or adjustment by regression. The Copas adjustment of treatment effects was not possible due to numerical issues in 143 (2.6%) meta-analyses.

We performed complete-case analysis with some missing data in the explanatory variables used in the mixed-effects models. For model 1, missing data were in publication year (14%), journal rank (31%) and number of citations (32%) with a total of 44% of the cases excluded. The missing data were due to published studies in journals with no journal rank or no citation information available. Google Scholar only provides data if the study has at least a single citation. For model 2, missing data were in the median journal rank (6%) and median number of citations (11%) with a total of 14% of the cases excluded. For models 3–6, missing data were in the median journal rank (2%), and median number of citations (3%) with a total of 4% of the cases excluded.

Changes from the initial protocol

The following six changes from the initial protocol were made. (1) We used a different categorisation into medical specialties. According to the protocol we aimed to use the 36 topics, but after acquiring the data we found that systematic reviews can be assigned to multiple such topics. Therefore, we used the 53 Cochrane Review Groups with every review assigned to exactly one review group. (2) We used the binary outcome of whether the effects were statistically significant but did not specify this clearly and just used the term 'effect size' instead of 'statistical significance of effect size' (protocol section 3: 1c and 2c; 4.3: 1c and 2c). (3) We specified the regression model for hypothesis 2c but we did not list the dependent variable in the statistical analysis plan (protocol section 4.3; 2c). (4) We were able to add a binary variable randomised controlled trial (RCT) (yes/no), that is, whether a study was reported as an RCT, and used this additional variable in the modelling (protocol section 4.3; 1c). (5) We used categorical quartiles of the journal rank instead of the continuous variable (protocol section 4.3; 1c); these were provided by Scimago themselves. (6) We applied Harbord's test only for ORs and not generally for all binary outcomes as specified in the protocol (protocol section 4.4).

Patient and public involvement

There was no patient and public involvement in this study.

RESULTS

Effect sizes from primary studies

Descriptives of primary studies

We analysed a total of 276 904 treatment effects from 57 162 studies that were published between 1922 and 2019. Online supplemental tables 3–5 give a detailed description of the primary studies per medical specialty (number of studies, median sample size, median treatment effect, etc).

Observed statistically significant effects in primary studies

Generally, we observed more statistically significant effects for efficacy outcomes than for adverse events with the exception of oncology and pregnancy and childbirth (figure 2A, online supplemental table 6). The proportions of significant effects in efficacy and safety outcomes differed among specialties (both $p<0.0001$). The largest proportion of significant treatment effects were 36% in both the specialties skin and wounds (95% CI 0.35 to 0.38) and anaesthesia and pain (95% CI 0.36 to 0.37); the smallest was 18% in neonatology (95% CI 0.18 to 0.19). For safety outcomes, the largest proportion was 29% in oncology (95% CI 0.27 to 0.31) and the smallest of 8% was in emergency and trauma (95% CI 0.07 to 0.10). We found no evidence for an association between the proportion of significant effects and the median sample size across the medical specialties (figure 2B). We performed a p curve analysis and found that all the specialties had a right-skewed shape, and no inflation of p values close to p=0.05 (online supplemental figure 2).

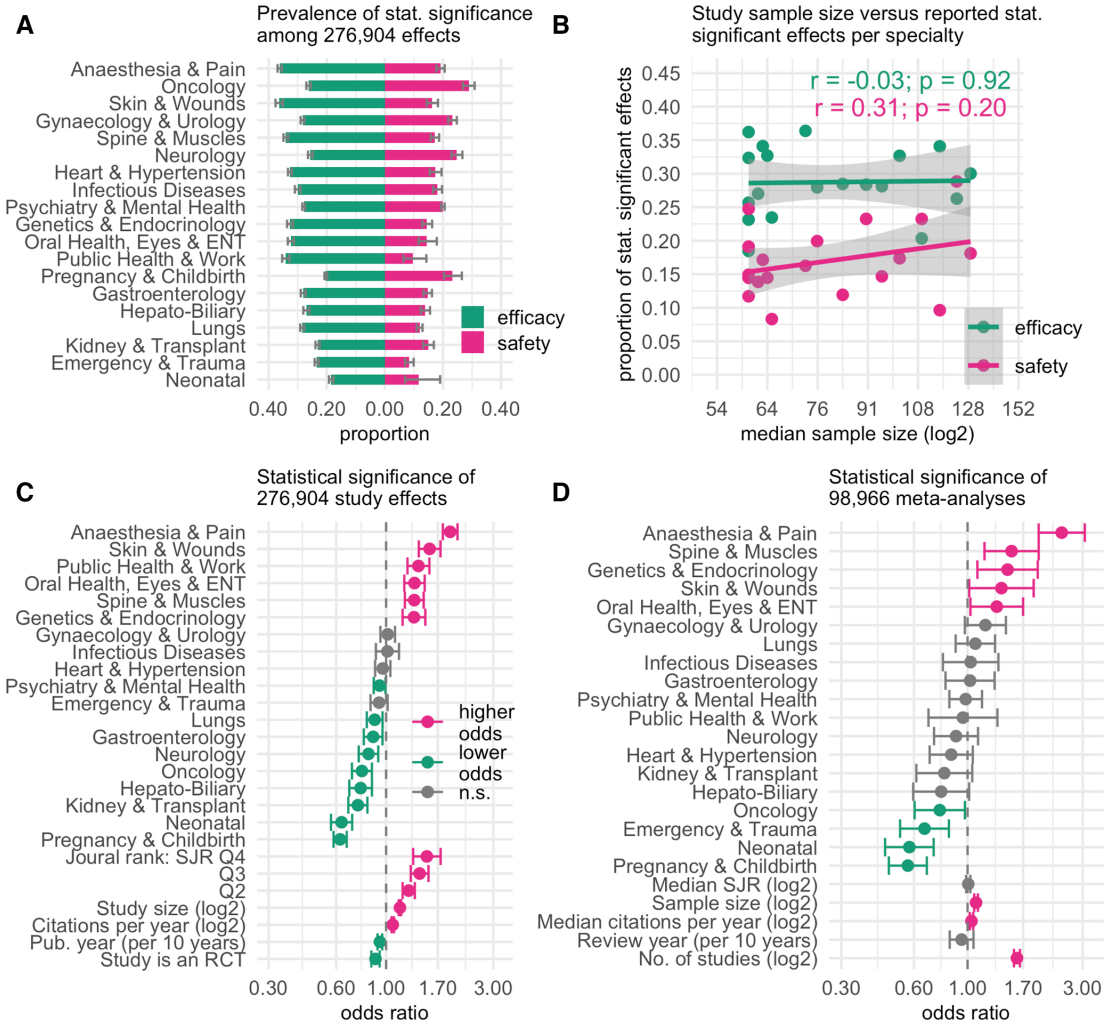


Figure 2 Reported statistically significant results across medical specialties. (A) Proportion (with 95% CI) of reported statistically significant effects in published studies for efficacy and safety outcomes. (B) Relationship between the proportion of statistically significant effects and the median sample size. ORs (with 95% CI) for a statistically significant effect in a primary study (C) and in a combined effect from a meta-analysis (D). RCT, randomised controlled trial. ENT, ear, nose and throat. SJR, Scimago journal rank. n.s., not statistically significant.

Odds for a statistically significant effect reported in a primary study

We modelled the binary outcome of whether a reported treatment effect was statistically significant. We found that the odds for a significant result was highest with a 1.9-fold increase in anaesthesia and pain (OR=1.93; 95% CI 1.79 to 2.08), and lowest with a 38% reduction in pregnancy and childbirth (OR=0.62; 95% CI 0.58 to 0.67) compared with the mean across specialties, see figure 2C and online supplemental table 7. Primary studies published in the second, third or fourth journal rank quartiles were all associated with higher odds for a statistically significant result (compared with the first quartile); the strongest effect had the top quartile (Q4) that showed a 1.5-fold increase of the odds (OR=1.52; 95% CI 1.32 to 1.75). Positive associations with significance were found for $\log_2$ study size (OR=1.15; 95% CI 1.14 to 1.17) and $\log_2$ number of citations per year (OR=1.07; 95% CI 1.06 to 1.08) and negative associations for publication year (per 10 years; OR=0.94; 95% CI 0.92 to 0.96) and when a primary study was an RCT (OR=0.90; 95% CI 0.86 to 0.94).

Combined effect estimates from meta-analyses

Descriptives of the meta-analyses

A total of 98 966 random effects meta-analyses were calculated from 4737 intervention reviews. The largest number of meta-analyses were from psychiatry and mental health (13 906); and the median total sample size in the meta-analyses across studies was N=523 (IQR 220–1395), see more descriptives in online supplemental table 8.

Observed statistically significant combined effects in meta-analyses

The proportion of statistically significant results varied across specialties (p<0.0001), the largest was 49% in anaesthesia and pain (95% CI 0.47 to 0.50), and the lowest was 21% in neonatology (95% CI 0.20 to 0.22), see online supplemental table 9. We performed a p curve analysis

and observed that all the specialties had a right-skewed shape with no inflation of p values close to p=0.05 (online supplemental figure 3).

Odds for a statistically significant effect in a meta-analysis

We modelled the binary outcome of statistical significance of the combined effects from the meta-analyses (figure 2D and online supplemental table 10). We found the highest odds with a 2.5-fold increase in anaesthesia and pain (OR=2.46; 95% CI 1.97 to 3.06) and the lowest with a 43% reduction in pregnancy and childbirth (OR=0.57; 95% CI 0.47 to 0.68) compared with the average across specialties. The number of studies (log_2) included in a meta-analysis had 1.6-fold higher odds for a statistically significant result (OR=1.60; 95% CI 1.56 to 1.65). Positive associations with significance had the log_2 sample size (OR=1.08; 95% CI 1.07 to 1.10) and the log_2 number of citations (OR=1.04; 95% CI 1.02 to 1.06). Journal rank and review year showed no evidence for an effect.

Small-study effects and publication bias

Prevalence of small-study effects

A total of 5534 meta-analyses from 949 Cochrane systematic reviews were assessed that satisfied our inclusion criteria. We found small-study effects in 1054 (19%) of the meta-analyses (95% CI 18% to 20%) based on Egger's and Harbord's tests, the distribution of p values is given in figure 3A. Infectious diseases had the highest prevalence

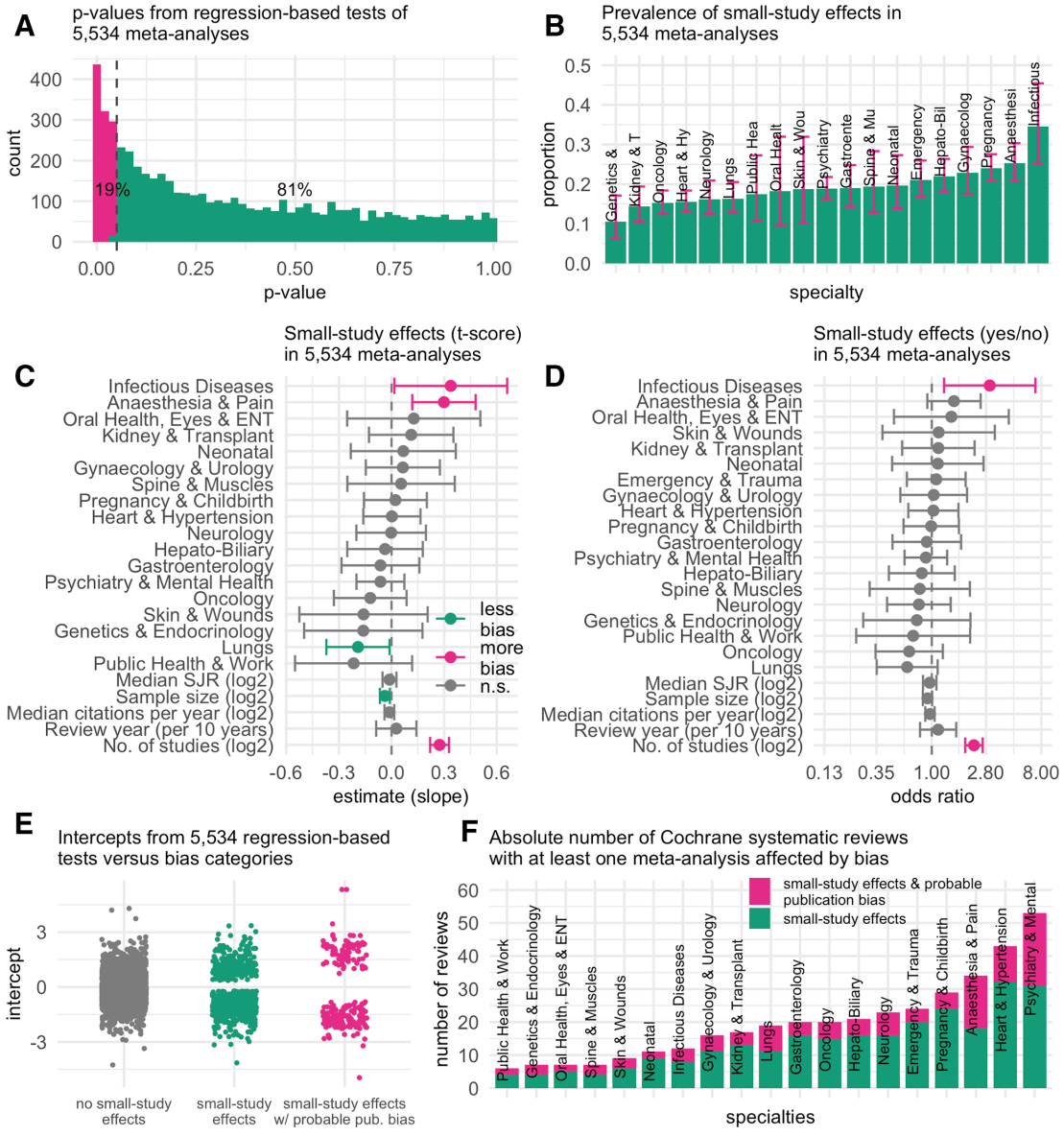


Figure 3 Small-study effects in meta-analyses across medical specialties. (A) Distribution of p values from regression-based tests, p<0.05 (one sided) were considered as evidence for small-study effects in meta-analyses (red). (B) Prevalence (with 95% CI) of small-study effects across the medical specialties. (C) Regression estimates (with 95% CI) for small-study effects with the outcome t-scores from regression-based tests. (D) Same but for a binary outcome (small-study effects yes/no). (E) Distribution of intercepts (magnitude of small-study effects) from regression-based test. (F) Number of systematic reviews affected by bias. ENT, ear, nose and throat. SJR, Scimago journal rank. n.s., not statistically significant.

(35%; 95% CI 25% to 45%) and genetics and endocrinology the lowest prevalence (10%; 95% CI 6% to 17%), see figure 3B.

Modelling small-study effects

We modelled the test statistic (t-score) from the regression-based test for small-study effects (figure 3C and online supplemental table 11). Infectious diseases (0.34; 95% CI 0.02 to 0.66) and anaesthesia and pain (0.30; 95% CI 0.12 to 0.48) showed larger small-study effects compared with the mean across the specialties; lungs showed smaller small-study effects (−0.19; 95% CI −0.37 to −0.01). The number of studies ($\log_2$) (0.27; 95% CI 0.22 to 0.33) was associated with larger small-study effects, the sample size ($\log_2$) with smaller (−0.04; 95% CI −0.07 to −0.01).

Similar results were obtained for small-study effect as binary outcome. Infectious diseases had threefold increased odds for small-study effects compared with the mean across specialties (OR=3.00; 95% CI 1.26 to 7.16), see figure 3D and online supplemental table 12. The number of studies ($\log_2$) included in the meta-analyses had twofold higher odds for small-study effects (OR=2.23; 95% CI 1.89 to 2.63).

Small-study effects with probable publication bias

Meta-analyses with evidence for small-study effects were further assessed. From the 1054 meta-analyses with small-study effects we identified 214 meta-analyses (20%) having small-study effects with probable publication bias which was 3.9% (95% CI 3.4% to 4.4%) of the total of 5534 meta-analyses included. The intercepts from Egger's tests (representing the magnitude of the bias) showed more extreme values for meta-analyses with probable publication bias than without (figure 3E); 99.5% of these were considered as substantial small-study effects based on the categorisation by Lin *et al*.[25]

Overall, 378 (40%) from 949 Cochrane systematic reviews had at least one meta-analysis with small-study effects and 115 (12%) reviews had at least one meta-analysis that showed small-study effects with probable publication bias, see figure 3F for results per medical specialty. Examples of small-study effects versus small-study effects with probable publication bias are shown in figure 4A,B, respectively.

Adjustment of effect estimates from meta-analyses

Regression-based adjustment of the 1054 meta-analyses with small-study effects resulted in a reduction of the effects in 99% of the cases and the median change in the combined effect (Pearsons's r) was −0.09 (figure 5A). For the Copas selection model the reduction was in 77% of the cases and the median change was −0.02 (figure 5B).

We applied a mixed-model to all 5534 included meta-analyses to find associations with effect change after regression-based adjustment and Copas adjustment (figure 5C,D). For the limit meta-analysis based on Egger's test, larger adjustments were required for Infectious diseases (−0.05; 95% CI −0.08 to −0.02) and anaesthesia and pain (−0.02; 95% CI −0.040 to −0.001). Also, Copas required larger adjustments for infectious diseases (−0.02; 95% CI −0.027 to −0.006) and anaesthesia and pain (−0.01; 95% CI −0.016 to −0.003). Kidney and transplant had smaller adjustments; but this was weak evidence. Furthermore, the sample size ($\log_2$) was associated with smaller adjustments with both methods, the median journal rank (SJR) with smaller adjustment with Copas, and the number of studies was associated with more adjustment for regression, see online supplemental tables 13 and 14.

Adjustment by regression generally performed a stronger shrinkage of the treatment effects compared with the Copas selection model analysis. However, the adjustments were consistent between the two methods across the medical specialties and across meta-analyses (figure 5E,F). Compared with the regression-based method, the Copas method applied more zero or near-zero adjustment.

Change of evidence for treatment effects

The change of evidence via Bayes factor bounds in 5534 meta-analyses was assessed by comparing the strength of the evidence after adjustment. The adjustment of the effects changed the overall evidence in many meta-analyses when effects were adjusted with regression but not so with Copas. The proportion of meta-analyses with very strong or decisive evidence for a treatment effect decreased from 32% down to 7% (regression) but only from 32% to 31% for Copas adjustments, see figure 5G.

DISCUSSION

Our study found that 19% from 5534 meta-analyses had evidence for small-study effects. This demonstrated that smaller studies reported larger effects, affecting 40% of the Cochrane systematic reviews. Most evidence for small-study effects was found in infectious diseases which also required larger adjustment compared with other specialties. Unsurprisingly, the number of studies in a meta-analysis was strongly associated with small-study effects as a higher number of studies increases statistical power to detect bias. Our results were consistent with the previously reported prevalence of asymmetry in 19% from a total of 366 meta-analyses[15] with exactly the same inclusion criteria. However, our dataset was 15 times larger as we also considered continuous outcomes which suggests a broader generalisability of our results.

Systematic assessment of meta-analyses with small-study effects suggested that publication bias may not be the main driver behind small-study effects. Only about 4% of the 5534 meta-analyses demonstrated small-study effects with probable publication bias. It is worth noting that regression-based methods do not take into account the statistical significance of the study effects. Therefore, other reasons for smaller studies reporting larger effects must be considered: confirmation bias, lower-quality studies, selection bias in the treatment groups, outcome

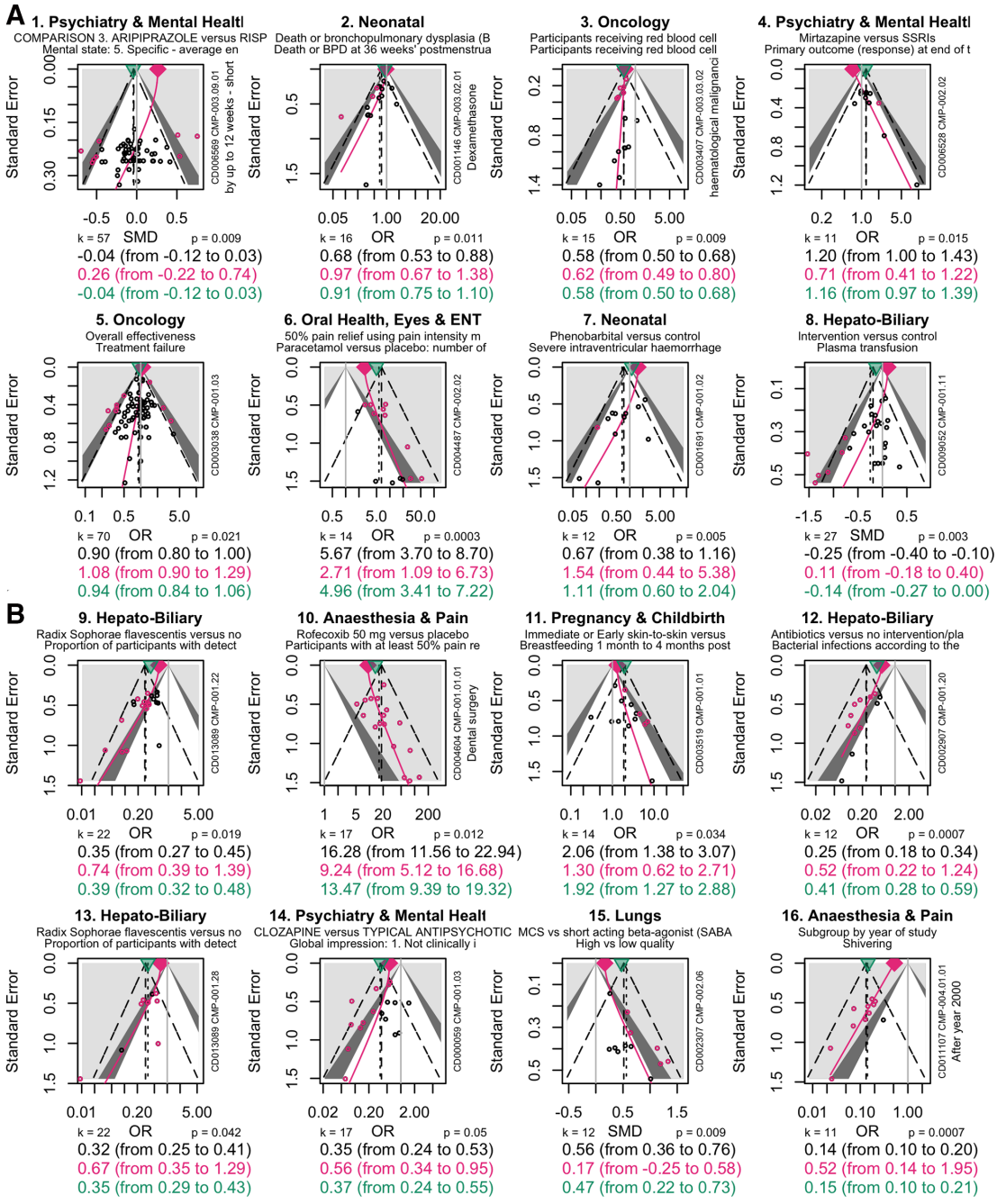


Figure 4 Enhanced funnel plots. Randomly picked examples of funnel plots with small-study effects (A); and small-study effects with probable publication bias (B). Small black circles are studies; statistically significant ones in red (based on original effect measure). The shades of grey are different levels of statistical significance (grey is p<0.05, light grey is p<0.01). The grey vertical line is the null, the dashed line the fixed effects, the dotted line the random effects estimate. The red diamond (regression) and green rectangle (Copas) are the adjusted effects, also as numbers (with 95% CI) at the bottom (unadjusted random effects estimate in black). The p value is from the asymmetry test by regression, k the number of studies. Above the funnel plots are the medical specialty, the outcome and comparison name; on the right the review id, meta-analysis id and the subgroup name. SMD, standardised mean difference. MCS, mast cell stabilisers.

switching or post hoc searches for statistical significance along with the incomplete reporting of statistically non-significant results.[42] However, all these biases have the same effect on meta-analyses: they all lead to an exaggeration of the combined treatment effect.

Regression-based adjustment of meta-analyses changed the evidence of the treatment effects in many cases from decisive to weak while the Copas adjustment was more conservative with a majority of meta-analyses that received no or near zero adjustment. The two methods are fundamentally different: while regression is based on small-study effects, Copas analysis directly models publication bias and the selection process. Regression-based methods tend to overcorrect in some situations, for example,

Figure 5 Adjustment of treatment effects. (A) Distribution of effect changes (Pearson's r) from meta-analyses after adjustment by regression; (B) same for Copas adjustment. (C) Variables associated with effect changes, negative estimates reflect shrinkage of effects; (D) same for Copas adjustment. (E) Adjustment by regression versus Copas across specialties; (F) same across all meta-analysis. (G) Strengths of evidence against the null hypothesis of no treatment effect before (random effects meta-analysis) and after adjustment (regression and Copas). ENT, ear, nose and throat. SJR, Scimago journal rank. n.s., not statistically significant. MA, meta-analysis.

when the adjusted effect turned out to be smaller than the smallest effect reported in a study. Future studies may apply shrinkage methods[43] to regression-based adjustments.

We found that some medical specialties were more likely to report a statistically significant effect, for example, in anaesthesia and pain and skin and wounds while others were less likely to do so, for example, pregnancy and childbirth and neonatology. These findings were confirmed when looking at combined effects from meta-analyses. Reasons for this could be different study types and endpoints across the specialties, a higher number of true effects in a field, varying statistical power between fields or differences in reporting of statistically significant

versus non-significant results. We replicated previous findings that studies with a statistically significant effect were more likely to be published in high ranking journals[8] and also received a higher number of citations[9 44] which can indeed contribute to publication bias.[45] We also showed that studies that are supposed to provide stronger evidence, such as RCTs, were less likely to report a statistically significant result.

Our study had some limitations. First, regression-based methods have been criticised to have a lack of statistical power[46] and there are also other situations where regression methods (and also selection models) may fail: large heterogeneity, no significant studies or all studies are of similar sizes. Therefore, we followed available guidelines for inclusion criteria of suitable meta-analyses,[14 15] among them including meta-analysis with low heterogeneity (I^2 <50%) only. However, this can be criticised as selection bias can be introduced. For example, it has been shown it may be favouring small studies in meta-analyses.[47] Furthermore, I^2 is not an absolute measure of heterogeneity, that is, it does not tell us how much the effects actually vary.[48] Nevertheless, the application of I^2 in our situation (to compare the amount of dispersion among studies within the CDSR) does not appear to be too unorthodox.[49] Next, we only included 5534 (5.6%) of the 98 966 meta-analyses calculated; most were excluded because the number of studies was smaller than ten. However, this selection of meta-analyses may represent the more relevant medical interventions where many studies have been conducted. At the same time, it does not mean that the excluded meta-analyses with less than 10 studies are of less importance as many highly relevant interventions in disease prevention and public health remain under-researched. In some cases, our meta-analyses do not exactly correspond to the ones performed in Cochrane Systematic Reviews as these also include unpublished findings; we excluded these as our aim was to assess publication bias when unpublished data is not considered. Altogether, systematic reviewers considered unpublished studies in 115 (11%) of the 1054 meta-analyses with small-study effects, and in 26 (12%) of the 214 meta-analyses that have small-study effects with probable publication bias. A strength of our study was that we carefully labelled efficacy versus adverse event outcomes, however, we used an automatic approach by matching keywords which depended on the correct documentation of the comparison and outcome name by the systematic reviewers. The grouping of specialties is a limitation in our study as the grouping could have been done differently.

Around half of systematic reviews, in some specialties even less, do not evaluate publication bias and should do so more often.[50 51] However, our findings suggest that large, potentially exaggerated effects from small studies give rise to greater concern than publication bias. The implications of this study are as follows: First, systematic reviewers should use advanced funnel plots with contours,[52] regression tests or Copas selection models to investigate the robustness of their conclusions. Second, investigators should carefully assess the risk of bias and also provide individual patient data to combat publication bias.[53 54] Finally, clinicians and patients should be cautious of small studies, especially when they report larger effects than larger studies, and to be attentive to sections on risk of reporting and publication bias in systematic reviews.

Twitter Simon Schwab @SimonSchw4b

Acknowledgements We thank Maya B. Mathur for commenting on the protocol and Philip Heesen for commenting on an earlier draft of the manuscript. We thank Cochrane for providing the data and their tremendous efforts to synthesize evidence in health care.

Contributors SS and LH designed the study. SS developed software and collected the data. SS and GK analysed the data. SS wrote the first draft of the manuscript and all authors commented on subsequent versions.

Funding SS received funding from SfwF (Stiftung für wissenschaftliche Forschung an der Universität Zürich; grant no. STWF-19-007).

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data used in this study are publicly available from the Cochrane Library (https://www.cochranelibrary.com/), Google Scholar (https://scholar.google.com/) and SCImago (https://www.scimagojr.com/). All data used in this study are openly available from the Cochrane Library (https://www.cochranelibrary.com/), Google Scholar (https://scholar.google.com/) and SCImago (https://www.scimagojr.com/). Data are also available from the corresponding authors upon reasonable request.

ORCID iD

Simon Schwab http://orcid.org/0000-0002-1588-2689

REFERENCES

1 Rothstein HR, Sutton AJ, Borenstein M. *Publication bias in meta-analysis*. Chichester: Wiley, 2006.
2 Egger M, Smith GD. Bias in location and selection of studies. *BMJ* 1998;316:61–6.
3 Sterling TD. Publication decisions and their possible effects on inferences drawn from tests of Significance-Or vice versa. *J Am Stat Assoc* 1959;54:30–4.
4 Dwan K, Altman DG, Arnaiz JA, *et al*. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One* 2008;3:e3081.
5 Decullier E, Lhéritier V, Chapuis F. Fate of biomedical research protocols and publication bias in France: retrospective cohort study. *BMJ* 2005;331:19.
6 Song F, Parekh-Bhurke S, Hooper L, *et al*. Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. *BMC Med Res Methodol* 2009;9:79.
7 Dickersin K, Chan S, Chalmers TC, *et al*. Publication bias and clinical trials. *Control Clin Trials* 1987;8:343–53.
8 Easterbrook PJ, Berlin JA, Gopalan R, *et al*. Publication bias in clinical research. *Lancet* 1991;337:867–72.

9 Nieminen P, Rucker G, Miettunen J, *et al*. Statistically significant papers in psychiatry were cited more often than others. *J Clin Epidemiol* 2007;60:939–46.
10 DeVito NJ, Bacon S, Goldacre B. Compliance with legal requirement to report clinical trial results on ClinicalTrials.gov: a cohort study. *Lancet* 2020;395:361–9.
11 Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol* 2000;53:1119–29.
12 Egger M, Davey Smith G, Schneider M, *et al*. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629–34.
13 Lau J, Ioannidis JPA, Terrin N, *et al*. The case of the misleading funnel plot. *BMJ* 2006;333:597–600.
14 Sterne JAC, Sutton AJ, Ioannidis JPA, *et al*. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011;343:d4002.
15 Ioannidis JPA, Trikalinos TA. The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *CMAJ* 2007;176:1091–6.
16 Harbord RM, Egger M, Sterne JAC. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med* 2006;25:3443–57.
17 Higgins JPT, Thomas J, Chandler J. *Cochrane Handbook for systematic reviews of interventions*. John Wiley & Sons, 2019.
18 Copas JB, Shi JQ. A sensitivity analysis for publication bias in systematic reviews. *Stat Methods Med Res* 2001;10:251–65.
19 Mavridis D, Salanti G. How to assess publication bias: funnel plot, trim-and-fill method and selection models. *Evid Based Ment Health* 2014;17:30.
20 Carpenter JR, Schwarzer G, Rücker G, *et al*. Empirical evaluation showed that the Copas selection model provided a useful summary in 80% of meta-analyses. *J Clin Epidemiol* 2009;62:624–31.
21 Moreno SG, Sutton AJ, Turner EH, *et al*. Novel methods to deal with publication biases: secondary analysis of antidepressant trials in the FDA trial registry database and related Journal publications. *BMJ* 2009;339:b2981.
22 Sutton AJ, Duval SJ, Tweedie RL, *et al*. Empirical assessment of effect of publication bias on meta-analyses. *BMJ* 2000;320:1574–7.
23 van Aert RCM, Wicherts JM, van Assen MALM. Publication bias examined in meta-analyses from psychology and medicine: a meta-meta-analysis. *PLoS One* 2019;14:e0215052.
24 Kicinski M, Springate DA, Kontopantelis E. Publication bias in meta-analyses from the Cochrane database of systematic reviews. *Stat Med* 2015;34:2781–93.
25 Lin L, Shi L, Chu H, *et al*. The magnitude of small-study effects in the *Cochrane Database of Systematic Reviews*: an empirical study of nearly 30 000 meta-analyses. *BMJ Evid Based Med* 2020;25:27–32.
26 Lin L, Chu H, Murad MH, *et al*. Empirical comparison of publication bias tests in meta-analysis. *J Gen Intern Med* 2018;33:1260–7.
27 Shi L, Lin L. The trim-and-fill method for publication bias: practical guidelines and recommendations based on a large database of meta-analyses. *Medicine* 2019;98:e15987.
28 Hayashino Y, Noguchi Y, Fukui T. Systematic evaluation and comparison of statistical tests for publication bias. *J Epidemiol* 2005;15:235–43.
29 Moreno SG, Sutton AJ, Ades AE, *et al*. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Med Res Methodol* 2009;9:2.
30 Murad MH, Wang Z. Guidelines for reporting meta-epidemiological methodology research. *Evid Based Med* 2017;22:139–42.
31 von Elm E, Altman DG, Egger M, *et al*. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007;335:806–8.
32 Viechtbauer W. Conducting Meta-Analyses in *R* with the metafor Package. *J Stat Softw* 2010;36:1–48.
33 Cooper H, Hedges LV, Valentine JC. *The Handbook of research synthesis and meta-analysis*. 2nd edition. Russell Sage Foundation, 2009.
34 Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. *Evid Based Ment Health* 2019;22:153–60.
35 Mueller KF, Meerpohl JJ, Briel M, *et al*. Methods for detecting, quantifying, and adjusting for dissemination bias in meta-analysis are described. *J Clin Epidemiol* 2016;80:25–33.
36 Schwarzer G, Carpenter JR, Rücker G. metasens: advanced statistical methods to model and adjust for bias in meta-analysis, 2019. Available: https://CRAN.R-project.org/package=metasens
37 Rücker G, Schwarzer G, Carpenter JR, *et al*. Treatment-Effect estimates adjusted for small-study effects via a limit meta-analysis. *Biostatistics* 2011;12:122–42.
38 Rücker G, Carpenter JR, Schwarzer G. Detecting and adjusting for small-study effects in meta-analysis. *Biom J* 2011;53:351–68.
39 Sellke T, Bayarri MJ, Berger JO. Calibration of ρ Values for Testing Precise Null Hypotheses. *Am Stat* 2001;55:62–71.
40 Held L, Ott M. On *p* -Values and Bayes Factors. *Annu Rev Stat Appl* 2018;5:393–419.
41 Bates D, Mächler M, Bolker B, *et al*. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw* 2015;67:1–48.
42 Chan A-W, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ* 2005;330:753.
43 Copas JB. Using regression models for prediction: shrinkage and regression to the mean. *Stat Methods Med Res* 1997;6:167–83.
44 Unger JM, Barlow WE, Ramsey SD, *et al*. The scientific impact of positive and negative phase 3 cancer clinical trials. *JAMA Oncol* 2016;2:875–81.
45 De Oliveira GS, Chang R, Kendall MC, *et al*. Publication bias in the anesthesiology literature. *Anesth Analg* 2012;114:1042–8.
46 Furuya-Kanamori L, Xu C, Lin L, *et al*. P value-driven methods were underpowered to detect publication bias: analysis of cochrane review meta-analyses. *J Clin Epidemiol* 2020;118:86–92.
47 Rücker G, Schwarzer G, Carpenter JR, *et al*. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Med Res Methodol* 2008;8:79.
48 Borenstein M, Higgins JPT, Hedges LV, *et al*. Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Res Synth Methods* 2017;8:5–18.
49 Higgins JPT, Thompson SG, Deeks JJ, *et al*. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557–60.
50 Hedin RJ, Umberham BA, Detweiler BN, *et al*. Publication bias and Nonreporting found in majority of systematic reviews and meta-analyses in anesthesiology journals. *Anesth Analg* 2016;123:1018–25.
51 Herrmann D, Sinnett P, Holmes J, *et al*. Statistical controversies in clinical research: publication bias evaluations are not routinely conducted in clinical oncology systematic reviews. *Ann Oncol* 2017;28:931–7.
52 Peters JL, Sutton AJ, Jones DR, *et al*. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *J Clin Epidemiol* 2008;61:991–6.
53 Page MJ, Sterne JAC, Higgins JPT, *et al*. Investigating and dealing with publication bias and other reporting biases in meta-analyses of health research: a review. *Res Synth Methods* 2021;12:248–59.
54 Ahmed I, Sutton AJ, Riley RD. Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey. *BMJ* 2012;344:d7762.