Pitfalls in reporting sample size calculation in randomized controlled trials published in leading anaesthesia journals: a systematic review

M. Abdulatif*, A. Mukhtar and G. Obayah
Anaesthetic Department, Faculty of Medicine, Cairo University, Giza, Egypt

*Corresponding author: 1 Al-Saray Street, Al-Manial, Cairo Postal Code 11559, Egypt. E-mail: mohamedabdulatif53@gmail.com

Abstract
We have evaluated the pitfalls in reporting sample size calculation in randomized controlled trials (RCTs) published in the 10 highest impact factor anaesthesia journals.

Superiority RCTs published in 2013 were identified and checked for the basic components required for sample size calculation and replication. The difference between the reported and replicated sample size was estimated. The sources used for estimating the expected effect size ($\Delta$) were identified, and the difference between the expected and observed effect sizes ($\Delta$ gap) was estimated.

We enrolled 194 RCTs. Sample size calculation was reported in 91.7% of studies. Replication of sample size calculation was possible in 80.3% of studies. The original and replicated sample sizes were identical in 67.8% of studies. The difference between the replicated and reported sample sizes exceeded 10% in 28.7% of studies. The expected and observed effect sizes were comparable in RCTs with positive outcomes ($P=0.1$). Studies with negative outcome tended to overestimate the effect size ($\Delta$ gap 42%, 95% confidence interval 32–51%), $P<0.001$. Post hoc power of negative studies was 20.2% (95% confidence interval 13.4–27.1%). Studies using data derived from pilot studies for sample size calculation were associated with the smallest $\Delta$ gaps ($P=0.008$).

Sample size calculation is frequently reported in anaesthesia journals, but the details of basic elements for calculation are not consistently provided. In almost one-third of RCTs, the reported and replicated sample sizes were not identical and the assumptions for the expected effect size and variance were not supported by relevant literature or pilot studies.

Key words: research hypothesis, effect size; statistical power, sample size; study design, superiority trials

Editor’s key points

- In this systematic review, the authors identify errors in sample size calculation in the leading anaesthesia journals.
- Frequently, there were differences between the reported and replicated sample sizes, with the assumptions for the expected effect size not supported by relevant data.

A randomized controlled trial (RCT) is the most frequently used, valid method to evaluate clinical interventions in medical subspecialities.¹ The Consolidated Standards of Reporting Trials (CONSORT) group developed a set of evidence-based recommendations to improve the quality of design, analysis, and interpretation of RCTs.³ Sample size calculation contributes to the quality of RCTs⁵ and is an important item in the CONSORT checklist.⁶ Adequate reporting of sample size calculation should normally include four main components: the expected minimal clinically relevant difference between the study groups, the sd of measurements for continuous primary outcomes, the power of the study (generally set between 80 to 90%), and type I error (usually 5%).⁵ Charles and colleagues⁸ demonstrated that 43% of RCTs published in the six highest impact factor medical journals did not report all the required parameters for sample size calculation. Moreover, in 30% of the reports that gave enough data to
recalculate the sample size, the variation in the replicated sample size was greater than 10%. Major deficiencies in sample size calculations were also reported by other specialties. Previous reports indicated significant improvement in the frequency of reporting of sample size calculation from 52% to 86% in anaesthesia journals. However, the integrity of sample size reporting for RCTs published in anaesthesia literature was not previously evaluated. This systematic review evaluated the pitfalls in reporting sample size calculation in parallel-group superiority RCTs published in the 10 highest impact factor anaesthesia journals during a 1 year period.

Methods

The Medline database was searched via ‘Pubmed’ using the search terms ‘randomized controlled trials’, ‘controlled clinical trial’, and ‘randomized controlled trial’ for the period from January to December 2013. The tables of contents and the abstracts of clinical reports published in all issues of 10 of the leading anaesthesia journals were also screened to identify the relevant RCTs (Fig. 1). The choice of the included journals was based on their impact factor ranking at the time of conceptualization of this systematic review. All superiority RCTs with parallel-group designs were included in the analysis. Randomized controlled trials with crossover or factorial design, those involving animals, and simulation or manikin studies were excluded.

Two of the authors (M.A. and A.M.) independently extracted data from the full text of the selected articles relating to the design and sample size calculation in duplicate. Any disagreement was resolved by discussion. We created a checklist to assess the frequency and adequacy of reporting of sample size estimation (Table 1). Sample size calculation was considered to be done if the authors explicitly stated this in the methodology section. Furthermore, we checked for the availability of a clearly specified primary outcome measure, the basic components of sample size calculations, and the source for estimating the minimal expected effect size. An effect size (Δ) is a standardized measure that describes the magnitude of the difference between study groups. The effect size was calculated using formulae outlined in Appendix I according to the type of primary outcome and the number of the groups in each RCT (Appendix I).

The number of patients randomized and the observed effect sizes were extracted from the results section. We also noted

---

Fig 1 Flow chart of the included and excluded studies and details of sample size replication. *Journals are listed in alphabetical order. †The authors of this review contributed to this excluded study.
whether the results of the trial were statistically significant for the primary outcome. The reviewers were not blinded to the journal name and authors.

Sample size replication was primarily done using the same software and statistical test specified by the authors. In studies not reporting sample size estimation software, the Power Analysis and Sample Size (PASS 13 software; NCSS, LLC, Kaysville, UT, USA) was used for replication. Replicate estimations were done for studies providing all the required components for sample size estimation. For RCTs missing the details of the statistical test, we used the formulae adapted for $\chi^2$ test and two-tailed Student’s $t$-test for binary and continuous end points, respectively.\textsuperscript{19–21} When adjustments for multiple testing were not reported, we assumed no adjustment had been applied. Sample size replication was not possible for studies that did not report the expected effect size.

The difference between the reported and replicated sample size was defined as the reported sample size minus the recalculated sample size divided by the reported sample size.\textsuperscript{5} A difference of 10% or more indicated important deviation from the reported sample size\textsuperscript{7} and was confirmed by expert statistical consultation.\textsuperscript{6,8} The magnitude of the expected and observed dropout rates was recorded.

The difference between the expected minimal effect size and the observed effect size in the primary outcome was considered the $\Delta$ gap.\textsuperscript{22} The magnitude of the $\Delta$ gap was compared between studies with positive and negative outcomes. Furthermore, the impact of the source for assumption of the expected effect size on the $\Delta$ gap was evaluated. Post hoc power calculation was performed using standardized formulae for each study that reported a non-significant difference in the primary outcome.\textsuperscript{23}

### Statistical analysis

Continuous variables are expressed as means (SD) or medians (interquartile range; IQR) and discrete variables as counts (percentage), unless otherwise stated. We calculated the expected and observed $\Delta$ values and dropout rates across all trials and compared them using Student’s paired $t$-test. Differences among the groups were analysed using Mann–Whitney U-test or Kruskal–Wallis test, as appropriate. The software SPSS v15.0 for Windows (SPSS, Inc., Chicago, IL, USA) was used for statistical analysis.

### Results

A total 231 RCTs were identified during literature search. Of these reports, 194 studies were enrolled in the analysis and 37 did not fulfil the prespecified inclusion criteria.\textsuperscript{14} Sixteen studies (8.3%) did not report any calculation for sample size, while 178 (91.7%) reported one or more of the essential parameters of sample size estimation (Fig. 1). Most of the included studies did not specify the statistical test and the use of one-tailed or two-tailed hypothesis for sample size calculation. The $\alpha$ error was most frequently set at 5%, and the reported power of the studies ranged between 70 and 95%. Thirty-four of the enrolled studies included more than two groups, but only three studies used Bonferroni correction to adjust the level of $\alpha$ error to compensate for multiple comparisons (Table 2). Replication of sample size calculation was possible in 143 (80.3%) studies. In 35 (19.7%) studies, replication

---

**Table 1 Data extraction checklist**

- Primary outcome measure clearly specified
- Type of primary outcome measure
- Sample size estimation was described in the methodology section
- Source for justification and magnitude of the expected effect size ($\Delta$)
- Power of the study and $\alpha$ level
- The expected and reported dropout rates
- Number of study groups
- Type of statistical test used for sample size calculation
- Hypothesis testing: one-tailed, two-tailed, or unspecified
- Compensation for multiple comparison in studies including more than two groups
- Software used for sample size calculation
- The final number of patients analysed in each of the study groups
- Difference between reported and replicated sample size
- Results of the primary outcome: positive vs negative

---

**Table 2 Reported parameters required for sample size calculation in the 178 included studies. Values are expressed as number (percentage)**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Number of studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected effect size in studies with continuous primary outcome</td>
<td></td>
</tr>
<tr>
<td>Reported</td>
<td>111 (62.4%)</td>
</tr>
<tr>
<td>Not reported</td>
<td>4 (2.2%)</td>
</tr>
<tr>
<td>Expected effect size in studies with binary primary outcome</td>
<td></td>
</tr>
<tr>
<td>Reported</td>
<td>56 (31.5%)</td>
</tr>
<tr>
<td>Not reported</td>
<td>7 (3.9%)</td>
</tr>
<tr>
<td>Expected variance in studies with continuous primary outcome</td>
<td></td>
</tr>
<tr>
<td>Reported</td>
<td>87 (48.9%)</td>
</tr>
<tr>
<td>Not reported</td>
<td>28 (15.7%)</td>
</tr>
<tr>
<td>Source for justification of the expected effect size</td>
<td></td>
</tr>
<tr>
<td>Relevant published studies</td>
<td>92 (51.7%)</td>
</tr>
<tr>
<td>Pilot study</td>
<td>29 (16.3%)</td>
</tr>
<tr>
<td>Unspecified</td>
<td>57 (32.0%)</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td></td>
</tr>
<tr>
<td>Two tailed</td>
<td>50 (28.1%)</td>
</tr>
<tr>
<td>One tailed</td>
<td>1 (0.6%)</td>
</tr>
<tr>
<td>Unspecified</td>
<td>127 (71.3%)</td>
</tr>
<tr>
<td>Type I error ($\alpha$)</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>173 (97.2%)</td>
</tr>
<tr>
<td>0.01</td>
<td>5 (2.8%)</td>
</tr>
<tr>
<td>Number of studies including more than two groups</td>
<td>36 (20.2%)</td>
</tr>
<tr>
<td>Number of studies that compensated for multiple comparisons</td>
<td></td>
</tr>
<tr>
<td>Power of study</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>1 (0.6%)</td>
</tr>
<tr>
<td>0.8</td>
<td>131 (73.6%)</td>
</tr>
<tr>
<td>0.85</td>
<td>3 (1.7%)</td>
</tr>
<tr>
<td>0.87</td>
<td>1 (0.6%)</td>
</tr>
<tr>
<td>0.9</td>
<td>37 (20.7%)</td>
</tr>
<tr>
<td>0.95</td>
<td>5 (2.8%)</td>
</tr>
</tbody>
</table>
was not possible because of unspecified expected effect size (Fig. 1).

Our analysis indicated that the replicated and reported sample sizes were different in 46 (32.2%) studies. This difference exceeded 10% in 41 (28.7%) studies. Ninety-two studies reported a compensation for a possible dropout rate. However, the observed dropout rate at the conclusion of the study was significantly lower than the expected rate (Table 3).

The median expected effect sizes for dichotomous and continuous outcomes were 0.5 (IQR 0.4–0.63) and 0.8 (IQR 0.6–1.0), respectively. The median observed effect sizes for dichotomous and continuous outcomes were 0.42 (IQR 0.2–0.6) and 0.7 (IQR 0.3–1.0), respectively. The overall median Δ gaps for dichotomous and continuous outcomes were 9% (IQR −5 to 30%) and 9% (IQR −30 to 30%), respectively. In studies with positive outcome (73.7%), the expected and observed effect sizes were not significantly different [Δ gap −6.0%, 95% confidence interval (CI) −14 to 2.0%, P=0.1. In the studies with negative outcomes (26.3%), the expected effect size was significantly higher than the observed effect size (Δ gap 42%, 95% CI 32–51%), P<0.001. The mean (95% CI) post hoc power of studies with negative outcome was 20.2% (13.4–27.1%).

The assumption for the expected difference and variance between the control and intervention groups was based on relevant published reports in 92 (51.7%) studies and on pilot studies in 29 (16.3%) studies, while 57 (32.0%) studies did not specify the source of the expected difference. The Δ gaps were smaller for RCTs using pilot studies for estimating the magnitude of the expected effect size (P=0.008; Fig. 2).

**Discussion**

The results of this systematic review indicated that sample size calculation is frequently reported, in 91.7% of RCTs published in leading anaesthesia journals. However, the details of the basic elements required for appropriate sample size calculation or replication are not consistently reported. In almost one-third of the enrolled RCTs, the variation in the replicated sample size was greater than 10% and the assumption for the expected minimal effect size was not supported by relevant published literature or pilot studies. Studies with a negative outcome were associated with significant overestimation of the expected effect size. In contrast, the least variations between the expected and observed effect sizes were encountered in RCTs using pilot studies as a source for estimating the minimal expected effect size.

In line with the findings of this systematic review, variations between the reported and replicated sample sizes have been reported at the protocol development stage,24 25 in trial registration databases,26 and after publication in high-impact medical journals.3 This finding is difficult to explain and could be related to the suboptimal cooperation between expert statisticians and clinical investigators. Involvement of a statistician in sample size calculation has been recommended15 27 and was used in the present systematic review. Using multivariate logistic regression, Koletsi and colleagues8 recently reported that the involvement of a methodologist in planning statistical analysis and trial design is associated with two-fold higher odds in adequate reporting of sample size calculation (odds ratio=1.97, 95% CI 1.10–3.53). However, the final outcome of sample size estimation cannot be guaranteed without a clear estimate and justification for the expected clinically important effect size by the investigator.28 29 The rationale for choosing a predetermined magnitude for the expected effect size was not adequately justified in RCTs included in this systematic review. A recent evaluation of research protocols submitted to UK ethics committees indicated that only 3% provided a reasoned explanation for the chosen effect size.24

The effect size is one of the most important indicators of clinical significance. Unfortunately, this fact is frequently overlooked, and many researchers mistakenly relate statistically significant outcomes with clinical relevance.30 The concept of clinical relevance in the primary outcome is identified in literature by what is called the minimal clinically important difference (MCID), defined as ‘the smallest change that patients perceive as beneficial or detrimental, and is useful to aid the clinical interpretation of health status data, particularly in response to intervention’.31–33 The use of the MCID in calculation of appropriate sample size for RCTs is expected to improve their clinical relevance and offers a threshold above which outcome is experienced as relevant by the patient.34

The concept of MCID was not addressed in the RCTs included in this systematic review, but the source used for estimating the expected effect size and variance was identified in 66% of studies.

**Table 3** Details of the reported and replicated sample sizes and dropout rates (143 studies). Values are expressed as number (percentage) or median [interquartile range (IQR)]. "The expected dropout rate was significantly higher than the observed rate (P=0.0001)

<table>
<thead>
<tr>
<th>Reported sample size</th>
<th>Not finally achieved</th>
<th>Studies compensating for possible dropouts</th>
<th>Expected dropout rate</th>
<th>Observed dropout rate</th>
<th>Replicated sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>26 (18.2%)</td>
<td>92 (64.3%)</td>
<td>16.8% (IQR 9.5–20.3%)</td>
<td>4.6% (IQR 0–6.6%)*</td>
<td>97 (67.8%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>29 (20.3%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17 (11.9%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>41 (28.0%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5 (3.5%)</td>
</tr>
</tbody>
</table>

![Fig 2 The impact of the source of prediction of the magnitude of effect size on the calculated Δ gaps (the difference between the observed and expected effect size). Box represents median observations (horizontal rule), with 25th and 75th percentiles of observed data (top and bottom of box). Whiskers indicate the maximal and minimal values. *Denotes significance relative to the other two groups, P<0.008.](image-url)
and was mostly derived from relevant published literature and pilot studies. Our finding is in line with a recent review by Clark and colleagues, who reported that only 59% of the protocols for pain and anaesthesia RCTs submitted to UK research ethics committees reported the source for the value of expected effect size. A recent comprehensive systematic review and survey identified seven potential methods for specifying the target expected effect size in sample size calculation, with wide variability in awareness and use among expert trialists in UK and Ireland. The use of pilot studies was recognized by 90% of responders as a source for identifying the minimal expected effect size required for sample size calculation.

Clinical trials using data derived from pilot studies for sample size calculation were associated with the smallest Δ gaps between the expected and observed effect sizes. However, pilot studies are generally underpowered and could yield an imprecise estimate of the difference in the population. The use of the upper limit of the 95% CI of the variance derived from internal pilot studies has been shown to improve the precision of sample size estimation. For RCTs using variance derived from an external pilot study or published literature, the use of the 60% upper confidence limit of σ2 has been recommended. An additional strategy to optimize sample size calculation is the use of a prespecified adaptive approach that allows recalculation of the sample size during the course of a clinical trial, with subsequent adjustments to the initially planned size. This approach is based on revision of the event rate, the variance, or the treatment effect.

The expected and observed effect sizes were significantly different only in studies with negative outcomes. This might indicate overestimation of the expected treatment effect in the intervention group. Consistent with our findings, Aberegg and colleagues investigated the possible bias in the design of RCTs reporting negative results in mortality in adult critical care settings. The mean expected and observed differences in mortality rates were 10.1% and 1.4%, respectively (P<0.0001), with a Δ gap of 8.7%. Two possible explanations were suggested to account for the increased Δ gap in negative trials: first, the investigators might adjust the expected effect size to reduce sample size; ‘sample size samba’ and second, the unrealistic optimism about the efficacy of treatment in the intervention study arm.

This systematic review elicited several factors that might be associated with unrecognized underpowered RCTs; these include: inappropriate estimation of the expected effect size, incorrect mathematical estimation of sample size, failure to achieve the minimal target sample size, and waiving of sample size calculation. An earlier study assessing the quality of RCTs published in anaesthesia literature reported that 72% of studies that had negative results did not consider the possibility of an underlying type II error and inadequate α priori sample size estimation. It has been suggested that many of the published anaesthesia studies are underpowered. This is supported by the results of our post hoc power calculation for RCTs with negative outcome. It should be noted, however, that combining the MCID with the 95% CI has been suggested as a valuable strategy for a more appropriate clinical interpretation of the outcome of negative RCTs. If the MCID lies within the 95% CI of the observed effect size, the treatment may be clinically effective regardless of the point of estimate.

Randomized controlled trials that are too small could be deceiving, either by missing sensibly direct treatment impacts that would be clinically paramount or by overestimating the effect of treatment and finding it significant merely by chance. Furthermore, it is generally considered unethical to enrol participants in underpowered trials with low potential to provide a reliable answer to the research question. The same ethical concern extends to exposing patients to unjustified risks in large clinical trials with overestimated sample size.

The dropout rate during the conduct of RCTs may complicate statistical analysis by introducing bias in the findings and reducing statistical power. Long-term follow-up has been associated with increased dropout rate. Apart from the relatively high dropout rate that might be encountered in chronic pain studies, RCTs evaluating anaesthetic interventions tend to have short follow-up periods. This could account for the relatively low (4.6%) observed attrition rate in the present systematic review. This finding could be used as a guide in optimizing sample size calculation for future RCTs in anaesthesia.

Appropriate sample size estimation commonly integrates scientific precision and available resources. There is currently an emerging approach incorporating study-specific optimal α and β errors that considers the clinically relevant effect size and the cost rather than the statistical significance in sample size estimation. The ultimate goals of this approach are to improve the feasibility of the study and to minimize the consequences of type I and II errors on clinical decision making. In most of the RCTs included in this review, the α and β errors were commonly set by convention at a level of 5 and 20%, respectively. Studies that compared more than two groups did not indicate that adjustment of α error was considered to compensate for multiple comparisons. A similar observation was reported in a recent review of the integrity of sample size determination in original research protocols. Anaesthesia trials frequently use a longitudinal design to assess the change in groups’ outcome at multiple time points. Multiple end points, multiple treatment arms, or interim analyses of the data require an error adjustment. It has been estimated that reducing the α error from 0.05 to 0.01 to compensate for multiple tests will almost double the minimal sample size at a study power of 90%. Furthermore, estimation of sample size among studies with repeated measures mandates identification of variance at each time measurement and the pattern of correlation among points of repeated measurements. Therefore, for a multi-arm study or a study with repeated measures, it would be useful to conduct a pilot study to provide sufficient details about the pattern of mean, mean difference, and the variances of response variable at each time point.

Inadequate description of sample size calculation continues to be reported despite the recommendations of the CONSORT and the Standard Protocol Items: Recommendations for Interventionsal Trials (SPIRIT) groups. Adherence to these recommendations has been reported to be suboptimal in anaesthesia, critical care, and other medical journals. Furthermore, the instruction and obligation to adhere to all the elements in the CONSORT checklist in journal author guides has been reported to be heterogeneous. We therefore developed a simple structured algorithm based on our results and on review of the literature to guide investigators to conduct and report sample size estimation appropriately (Fig. 3). We also suggest that the transparency could be improved by adopting an obligatory short checklist of all the basic elements of sample size calculation at the different stages of designing and reporting of RCTs. Our assumption is in line with the recommendation of the World Health Organization Minimal Registration Data Set and is supported by the finding that perfect 100% compliance was reported in obligatory data, while a variable adherence was observed in optional data in one of the most popular trial registration databases.
Our review has some limitations. First, we have restricted our reappraisal of sample size estimation to parallel-group superiority RCTs to ensure uniform conclusions; therefore, our findings cannot be extended to other types of RCTs or other study designs. Second, our literature search covered all issues of the 10 highest impact factor anaesthesia journals published in 1 year. Probably, a wider spectrum could have pointed out additional shortcomings in sample size calculation. However, we planned to have the most up-to-date status of this issue by including only RCTs published in 2013. Third, we extracted relevant sample size data from published RCTs. Comparisons with the original study protocols of registered trials were not considered. Registration of intervention clinical trials is increasingly required in anaesthesia\(^70\) and other medical specialties.\(^71\) However, at the time of preparation of this systematic review, some of the selected journals were not yet adopting a compulsory prior trial registration policy.\(^72\)

In conclusion, the results of this systematic review indicated that despite a high frequency (91.7%) of sample size reporting, some of the required basic assumptions for calculation were deficient or not supported by plausible reasoning in 19.7 and 32% of studies, respectively. This could explain the large differences between the design assumptions and the observed data. The continued suboptimal reporting of sample size estimation calls for more strict guidelines at the different stages of designing and reporting of RCTs. We suggest that the use of our proposed simple algorithmic approach at the design stage and an obligatory checklist in journals’ author guides could improve the integrity of reporting sample size calculation. A follow-up future evaluation will be required to validate this recommendation.

**Authors’ contributions**

M.A. and A.M. contributed to the design of the review, data extraction, analysis of the results, writing, revision, and final approval of the manuscript. G.O. contributed to tabulation of data, analysis, critical revision, and final approval of the manuscript.

**Acknowledgements**

We would like to acknowledge the support and advice of Professor Jeffrey S. Kane for data analysis and optimization of sample size replication.

**Declaration of interest**

M.A. is a member of the associate editorial board of the British Journal of Anaesthesia. A.M. and G.O. have no conflict of interest to declare.

**Funding**

Departmental resources.
References

4. Todd MM. Clinical research manuscripts in Anesthesiology. Anesthesiology 2001; 95: 1051–3
32. Kvien TK, Heiberg T, Hagen KB. Minimal clinically important improvement/difference (MCII/MCID) and patient acceptable symptom state (PASS): what do these concepts mean? Ann Rheum Dis 2007; 66 Suppl 3: iii40–1
Appendix I: Formułae used for effect size calculations:

Continuous primary outcome

- For studies including two groups, the effect size was calculated by taking the difference in means and dividing that number by the sd of the control group (Glass’s $d$).  
- For studies including three or more groups, the effect size was calculated as follows.

We first calculated Cohen’s $d$ as follows:

$$d = \frac{m_{\text{max}} - m_{\text{min}}}{\sigma}$$

where $m_{\text{max}}$ and $m_{\text{min}}$ are the highest and lowest means, respectively, and $\sigma$ is the sd within population.

Then $d$ was transformed to $f$ using the following formula (Cohen’s $f$):

$$f = d \sqrt{\frac{1}{2k}}$$

where $k$ is the number of population means.
Binary primary outcome

• For studies including two groups, the effect size was calculated using the following formula:

\[
\frac{(p_1 - p_2)}{\sqrt{P(1-P)}}
\]

where \(p_1\) and \(p_2\) are the proportions in the two groups and \(P = (p_1 + p_2)/2\) is the mean of the two values. \(^5\)

• For studies with more than two proportions, \(\chi^2\) statistics was used to estimate the effect size. \(^18\)

\[
\omega = \sqrt{\frac{\sum_i (O_{ik} - E_{ik})^2}{E_{ik}}}
\]

where \(\omega\) is the effect size, \(O_{ik}\) and \(E_{ik}\) denote the observed and expected proportion, respectively, \(m\) is the total number of cells and \(i\) the cell number.

*Handling editor: J. G. Hardman*