Journal of Clinical Epidemiology 106 (2019) 50−59

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Increased risks for random errors are common in outcomes graded as high certainty of evidence

Gerald Gartlehner[a,b,*], Barbara Nussbaumer-Streit[b], Gernot Wagner[b], Sheila Patel[a], Tammeka Swinson-Evans[a], Andreea Dobrescu[c], Christian Gluud[d]

[a]*RTI International, Research Triangle Park, NC, USA*
[b]*Cochrane Austria, Danube University Krems, Krems, Austria*
[c]*Genetics Department, Victor Babes University of Medicine and Pharmacy, Timisoara, Romania*
[d]*Copenhagen Trial Unit; Centre for Clinical Intervention Research, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark*

Accepted 10 October 2018; Published online 19 October 2018

Abstract

Objectives: The aim of article was to assess the risk for random errors in outcomes graded as high certainty of evidence (CoE).

Study Design and Setting: We randomly selected 100 Cochrane reviews with dichotomous outcomes rated as high CoE using Grading of Recommendations Assessment, Development, and Evaluation. To detect increased risks for random errors, two investigators independently conducted trial sequential analysis using conventional thresholds for type I ($\alpha = 0.05$) and type II ($\beta = 0.10$) errors. We dually regraded all outcomes with increased risks for random errors and conducted multivariate logistic regression analyses to determine predictors of increased risks for random errors.

Results: Overall, 38% (95% confidence interval: 28−47%) of high CoE outcomes had increased risks for random errors. Outcomes assessing harms were more frequently affected than outcomes assessing benefits (47% vs. 12%). Regrading of outcomes with increased random errors showed that 74% should have been downgraded based on current guidance. Regression analyses rendered small absolute risk differences ($P = 0.009$) and low number of events ($P = 0.001$) as significant predictors of increased risks for random errors.

Conclusion: Decisionmakers need to be aware that outcomes rated as high CoE often have increased risks for false-positive or false-negative findings. © 2018 Elsevier Inc. All rights reserved.

Keywords: Decision-making; GRADE; Random errors; Systematic reviews; Trial sequential analysis

1. Introduction

Grading of Recommendations Assessment, Development, and Evaluation (GRADE) [1] and closely related assessment systems such as that of the Evidence-Based Practice Center (EPC) program of the US Agency for Healthcare Research and Quality (AHRQ) [2] have become commonly used tools to convey the "certainty of evidence" (CoE; equivalent terms are "quality of evidence" or "strength of evidence") in systematic reviews. For decision-makers, such assessments are crucial because they convey the confidence that reviewers have in the correctness of results and in where effects lie relative to particular thresholds that are relevant for decision-making [3].

In a previous project, we assessed the predictive value of CoE grades [4]. We determined how reliably these grades can predict whether observed treatment effects will change as new studies are incorporated into an existing evidence base. The current approach only partly fulfilled qualities of predictive value. Specifically, more than 20% of outcomes graded as high CoE substantially changed in magnitude of effect as new studies were added [4]. Because high CoE, by definition, conveys strong confidence in the correctness of estimates, such a high proportion of changing estimates raises concerns for clinical and policy decision-making.

Conceivably, several factors might account for the limited predictive value of grades of CoE. Important ones include (1) lack of adherence of systematic reviewers to published GRADE (or other) guidance, (2) problems with

Conflict of interest: None of the authors reports any conflicts of interest with respect to the topic of this article.

* Corresponding author. RTI International, 3040 East Cornwallis Road, Research Triangle Park, NC 27709, USA. Tel.: +43 2732 893 2910; fax: +43 2732 893 4910.

E-mail address: ggartlehner@rti.org (G. Gartlehner).

https://doi.org/10.1016/j.jclinepi.2018.10.009

0895-4356/© 2018 Elsevier Inc. All rights reserved.

What is new?

Key findings

- Trial sequential analysis (TSA) detected that more than one-third of outcomes rated as high certainty of evidence (CoE) based on GRADE (Grading of Recommendations Assessment, Development, and Evaluation) have increased risks for type I or type II errors.
- Outcomes assessing harms were affected more frequently than outcomes assessing benefits.
- Lack of adherence to guidance documents, particularly when assessing imprecision, appears to be a major issue when grading CoE.
- Small absolute risk differences and low event rates are statistically significant predictors of high risk of random errors in outcomes rated as high CoE.

What this adds to what was known?

- Previous research indicated low predictive value of grades of CoE. This study assesses the predictive value of CoE and explores reasons for the low predictive value of high CoE grades.

What is the implication and what should change now?

- Decision-makers need to be aware that a substantial proportion of outcomes rated as high CoE using GRADE might have an increased risk for false-positive or false-negative findings. Working groups such as those of GRADE or the Agency for Healthcare Research and Quality Evidence-based Practice Center program need to consider how they can better address these increased risks of random error, particularly when assessing the imprecision domain. They should also reflect on how to reduce unwarranted variation in the use of GRADE. Producers of systematic reviews need to provide better quality assurance regarding grades of CoE. TSA may assist in grading imprecision.

operationalizing the GRADE approach, or (3) ambiguous statements in published guidance documents. Another important factor may be the conceptual approach to grading the CoE. The current approach may not adequately take into consideration the risks for random errors (type I [false-positive] or type II [false-negative]) when assessing imprecision. Random errors can be increased in meta-analyses that do not reach the "required information size" (equivalent terms are "optimal information size" or "meta-analytic sample size") [5]. The required information size indicates the number of participants necessary in a meta-analysis to ensure that type I and type II errors are not larger than prespecified (usually 5% for type I errors and 10–20% for type II errors). Meta-analyses that do not reach the required information size have an increased risk of over- or under-estimating a treatment effect and leading to spurious inferences [5–7].

Although GRADE and related approaches consider required information size (GRADE calls it "optimal information size"), they do not take into account between-study heterogeneity within a meta-analysis and multiple testing during updates of systematic reviews [2,8]. Methods research, however, demonstrates that these factors can increase the risk for random errors [5–7]. Simulation studies indicate that repeated significance testing during regular updates of systematic reviews can increase the risk for type I errors to 30% or higher [9,10]. Cochrane reviews, for example, are required to be updated at least every second year [11].

The goal of this research project was to identify factors that could be responsible for the limited predictive value of high CoE grades [4]. Specifically, we intended to determine whether an increased risk for type I or type II errors could be the reason for the low predictive value of high CoE grades.

2. Methods

Two specific research goals guided our project:

1. To determine the proportion of outcomes rated as high CoE that have increased risks for type I or type II errors.
2. To determine whether increased risks for random errors can be attributed to inappropriate grading (e.g., authors do not follow current guidance) or flaws in the conceptual approach of GRADE.

To achieve the first research goal, we tested a randomly selected sample of high CoE outcomes with trial sequential analysis (TSA) [6]. To accomplish the second research goal, we closely examined outcomes that TSA determined to have an increased risk for random errors. Table 1 defines terms commonly used in this article.

2.1. Trial sequential analysis

TSA is a statistical approach that can reveal insufficient information size and reduce spurious inferences from meta-analyses because of type I or type II errors [14]. This approach combines conventional meta-analyses with methods to calculate information sizes and adjust trial sequential monitoring boundaries for benefit, harm, and futility [12,14]. Factors that increase the risk for random errors include sparse data, multiple testing, between-study heterogeneity, or lack of required information size [12]. TSA establishes sequential monitoring boundaries using the Lan-DeMets α-spending function with O'Brien-Fleming type boundaries [12,14–16].

Table 1. Definitions of commonly used terms

Boundary of superiority (or inferiority): Adjusted threshold that ensures that the risk of a false-positive conclusion (type I error) is smaller than the prespecified α before a meta-analysis has surpassed its required information size (RIS) [12].

Boundary of futility (or no effect): Adjusted threshold that assures that the risk of a false-negative conclusion (type II error) is smaller than the prespecified β before a meta-analysis has surpassed its RIS [12].

Required information size: The required number of participants in a meta-analysis to ensure that the maximum type I error is no larger than α, and that the maximum type II error is no larger than β when testing for statistical significance [8]. Stated differently, it is the required number of participants required to reject or accept a certain intervention effect based on the chosen parameters.

Certainty of evidence (CoE): The degree of confidence that estimates are close to the true effect [2]. Equivalent terms are *quality of evidence* or *strength of evidence*.

High CoE: Assessors are confident that the estimate of effect lies close to the true effect, and that the findings are stable, that is, that another study would not change the conclusions [2].

Type I error (α): The incorrect rejection of a true null hypothesis. In other words, the detection of an effect where none exists. The significance level is the probability of making a type I error. By convention, α is usually set to 5%, implying that a 5% probability of incorrectly rejecting the null hypothesis is acceptable [13].

Type II error (β): The incorrect acceptance of the null hypothesis, that is, the failure to detect an existing effect. By convention, the acceptable risk for type II errors is between 10% and 20% [13].

When results in meta-analyses are statistically significant according to the naïve 95% confidence limits, TSA determines whether the findings can be attributed to an increased risk of a type I error (relative to a prespecified type I error, which is usually 5%) or to actual superiority (or inferiority, based on a prespecified threshold) of one intervention compared with another. When results are statistically nonsignificant, TSA assesses whether the lack of statistical significance could be attributed to an increased risk of type II error (relative to a prespecified type II error, which is usually 10−20%) or to an underlying equivalence of interventions (based on a prespecified, plausible relevant difference in effect).

For this project, we used TSA to explore increased risks of random errors in outcomes rated as high CoE. To adhere to the current definition [17], any outcome rated as high CoE should have robust TSA results that allow firm conclusions about the presence or absence of an effect. Outcomes rated as high CoE that TSA results do not support have an increased risk for type I or type II errors.

2.2. Calculating sample sizes

Previous research indicated that up to 49% of statistically significant meta-analyses were inconclusive when adjusted for risks for random errors [18]. For our sample size calculations, we assumed that at least 10% of outcomes rated as high CoE could be affected. Based on this proportion and 90% power, a sample size calculation for a one-sample test indicated the need for at least 85 high CoE outcomes.

2.3. Designing a sampling frame

To obtain a random sample, we allocated a unique random number to each of the 1,565 Cochrane and 34 AHRQ systematic reviews that had been published in 2015 or 2016. We ranked the publications by random numbers to avoid bias attributable to publication date. Following this order, we then assessed each review for eligibility until we had 100 eligible reviews that met the following criteria:

- The review included at least one dichotomous outcome rated as high CoE;
- The high CoE outcome was based on a meta-analysis of at least three randomized controlled trials (RCTs);
- The review presented sufficient data to replicate the meta-analysis.

To avoid correlation effects, we included only one outcome per review. We selected the first high CoE outcome in each review that the authors had presented. For computational reasons, we did not use continuous outcomes because thresholds for clinical relevance are often more difficult to determine. By chance, all randomly selected systematic reviews were Cochrane reviews.

2.4. Extracting data from Cochrane reviews

One author extracted general information about the body of evidence for each eligible outcome. This included the type of intervention, type of outcome (assessing benefits or harms), sample size, number of included studies, baseline event rate, and other relevant information. A second author checked extracted data for correctness. We received RevMan files for each included Cochrane review from the Cochrane Editorial Unit; this enabled us to import data directly into the TSA program. To ensure correctness of imported data, we reran meta-analyses in TSA and compared our results with those in included reviews.

2.5. Applying trial sequential analyses

Two investigators independently conducted all TSA analyses using version 0.9.5.10 beta (Copenhagen Trial Unit, Denmark). We resolved discrepancies by consensus discussion. To conduct meta-analyses of relative risks, we used random effects models (DerSimonian-Laird), except in one case [19] for which we used a fixed effect model because a random effects model did not seem appropriate. In this analysis, the random effects model rendered a substantially different effect estimate (and conclusion) than the fixed effects model that authors of the Cochrane reviews had used.

To determine the required information size, we used conventional parameters for error probabilities: type I, two-sided α = 0.05; type II, two-sided β = 0.10; and the diversity of the meta-analysis as the measure of the heterogeneity [7]. We estimated the proportion of participants with an event in the control group based on the meta-analysis of the respective Cochrane report.

2.6. Assessing concordance and discordance of trial sequential analysis results with high CoE ratings

TSA classifies sequential monitoring boundaries into boundaries of superiority (or inferiority) and boundaries of futility. Boundaries for superiority (or inferiority) can be interpreted as thresholds for “conclusions of effect”; by contrast, boundaries of futility can be viewed as thresholds for “conclusions of no effect.”

We assessed whether the cumulative Z-curve crossed the required information size or any of the monitoring boundaries; in such cases, we viewed results as firm evidence for an intervention effect or firm evidence for the absence of such an effect, which is concordant with high CoE grades.

Fig. 1 illustrates the concept of a TSA analysis. For simplicity, we illustrate a one-sided test when the required information size (adjusted for heterogeneity and multiple testing) is not fulfilled. The meta-analysis of the included trials is depicted as a cumulative Z-curve.

In Fig. 1, points A—D on the dotted lines of the Z-curves illustrate four possible scenarios.

2.6.1. Increased risk for type I error

The cumulative Z-curve crosses the limit of conventional statistical significance ($P = 0.05$), but it does not cross the boundary of superiority (scenario A in Fig. 1). Such a result indicates an increased risk for a type I error and is discordant with precision and a high CoE grade.

2.6.2. Firm evidence of effect

The cumulative Z-curve crosses the boundary of superiority (scenario B in Fig. 1). Such a result is concordant with precision and a high CoE grade.

2.6.3. Increased risk for type II error

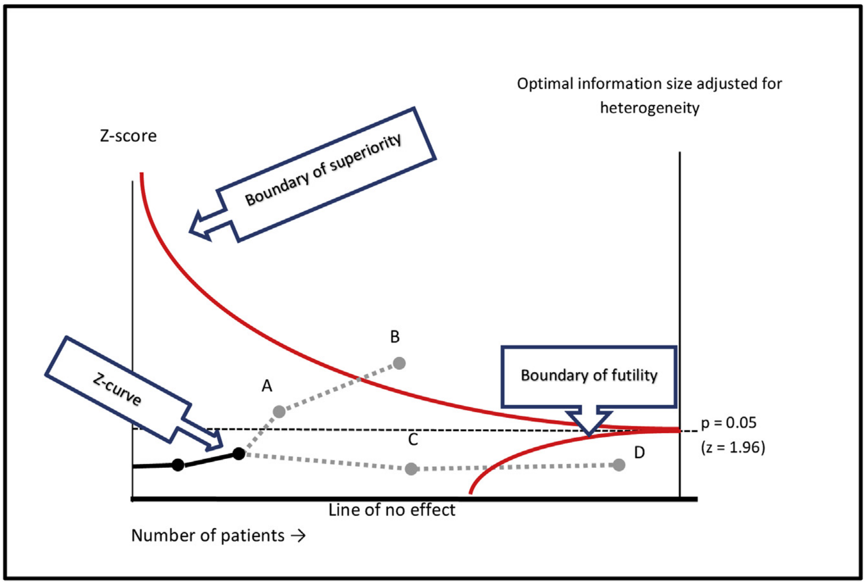
The cumulative Z-curve crosses neither the limit of conventional statistical significance nor the boundary of futility (scenario C in Fig. 1). Such a finding indicates an increased risk for a type II error and is discordant with precision and a high CoE grade.

2.6.4. Firm evidence of lack of effect

The cumulative Z-curve crosses the boundary of futility (scenario D in Fig. 1). Such a result is also concordant with precision and a high CoE grade.

For conclusions of no effect, we used three different intervention effects (10%, 15%, and 25% relative risk change) as margins of clinical relevance. When TSA did not yield consistent evidence of lack of effect for all three thresholds, two team members with clinical background provided a judgment about the clinical relevance of the relative risk reductions. For example, for patient-relevant outcomes such as mortality, pain, or hospital admission, we calculated absolute risk changes to determine whether 10% or 25% relative risk change should be viewed as a threshold for clinical relevance.

For conclusions of effect (benefits or harms), we used the observed treatment effect as a threshold to calculate monitoring boundaries.



* For illustrative purposes, the presentation is one-sided (for beneficial effects)

Fig. 1. Schematic of four scenarios of trial sequential analysis*. *For illustrative purposes, the presentation is one-sided (for beneficial effects).

For sensitivity analyses, we explored the impact of a 20% risk for type II error (instead of 10%). In addition, we determined whether using the effect estimate and the statistical model in the original analysis of the Cochrane review (e.g., odds ratio, fixed effect model) would change results. Finally, we omitted the TSA adjustment for multiple updates.

2.7. Exploring discordant results

To address the second research goal, we explored potential reasons for increased risks of random errors. To determine whether Cochrane authors adhered to current GRADE guidance, two team members with experience and formal training in GRADE independently rated the CoE for each discordant outcome. We resolved discrepancies by consensus or involvement of a senior team member. In cases in which two investigators agreed that an outcome should *not* be rated as high CoE, we explored possible reasons for nonadherence to guidance by Cochrane authors.

In addition, we conducted multivariate logistic regression analyses with backward elimination to detect factors that predict increased risks for random error. We used the Stata *logit* command to fit a logit model for a binary response by maximum likelihood. We modeled the probability of a discordant result given the following regressors: baseline event rate, conclusion of superiority or no effect, absolute risk difference, type of outcome (beneficial or harmful), number of participants in meta-analyses, number of studies in meta-analyses, number of events, type of intervention (e.g., pharmacologic, medical devices, surgical, and behavioral), and magnitude of effect. To assess the impact of collinearity on the variance of the parameter estimates, we calculated variance inflation factors. To estimate the goodness-of-fit of the model, we used the Pearson's goodness-of-fit test. We conducted descriptive and regression analyses with Stata, version 13.1 (StataCorp, TX, USA).

Table 2. Descriptive characteristics of included reviews

Type of interventions (%)	Pharmacologic: 70 Medical devices: 12 Surgical: 3 Educational: 3 Behavioral: 2 Infrastructure: 2 Physical: 2 Screening: 1 Complementary: 1 Complex interventions: 1 Radiotherapy: 1 Other: 2
Type of control interventions (%)	Inactive: 57 Active: 20 Usual care: 23
Type of outcomes (%)	Beneficial: 25 Harmful: 75
Number of included studies in meta-analyses: median (range)	6 (3–52)
Number of included participants in meta-analyses: median (range)	1,785 (37–43,290)
Number of events in meta-analyses: median (range)	316 (14–3,871)
Percentage of participants with event in control group: median (range)	16 (1–89)
Absolute risk difference (percentage points) between intervention and control groups: median (range) Outcomes with conclusions of effect Outcomes with conclusions of no effect	10 (1–58) 1 (0–5)
Relative risk difference (percentage) between intervention and control groups: median (range) Outcomes with conclusions of effect Outcomes with conclusions of no effect	44 (6–1,095) 7 (0–39)

3. Results

Overall, the included Cochrane systematic reviews provided data on more than 469,000 participants for the selected outcomes. On average (median), eligible meta-analyses included six RCTs with 1,785 participants; 65% reported statistically significant results.

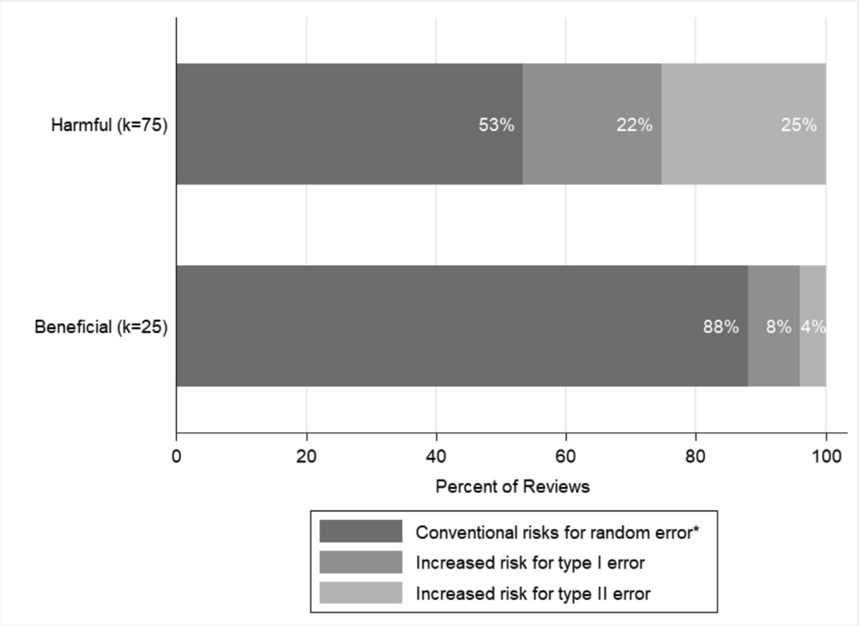
Table 2 summarizes descriptive characteristics of included reviews. Supplementary File 1 presents detailed characteristics of included reviews about populations, interventions, comparators, and outcomes.

3.1. Proportion of outcomes with increased risk for random error

In the following sections, we refer to risks for type I errors that are larger than 5% as "increased risks for type I errors" and risks for type II errors that are larger than 10% as "increased risks for type II errors." In all, we had 25 outcomes assessing benefits and 75 assessing harms.

In 13 instances involving conclusions of no effect, TSA resulted in inconsistent findings between 10% and 25% relative risk changes. Because a prespecified threshold is necessary for calculating the required information size in TSA, two investigators with clinical background determined which of the two thresholds would lead to a clinically relevant difference. In two cases (both on mortality), they viewed a 10% relative risk change as clinically relevant; in the remaining cases, they determined a 25% relative risk change as clinically relevant. We used their decisions for all further analyses.

Overall, 38% (95% confidence interval [CI]: 28–47%) of outcomes graded as high CoE in the Cochrane systematic reviews had increased risks for type I or type II errors when we adjusted for between-study heterogeneity, lack of required information size, and multiple testing. In other words, at least 28% but up to 47% of outcomes that Cochrane authors had rated as high CoE had higher than conventionally accepted risks of providing false-positive



* Conventional risks for random errors are defined as a risk for type I error of 5% and a risk for type II error of 10%; Abbreviations: k=number of studies

Fig. 2. Proportions of increased risks of random errors in outcomes assessing harms or benefits. *Conventional risks for random errors are defined as a risk for type I error of 5% and a risk for type II error of 10%; Abbreviations: k = number of studies.

or false-negative results. Supplementary File 2 presents the characteristics of meta-analyses of Cochrane reports with increased risks for random error.

Increased risks for type II errors were more common in our sample than increased risks for type I errors. Of 35 outcomes with a conclusion of no effect (no statistically significant difference), 20 (57%; 95% CI 40–73%) had an increased risk for a type II error. By comparison, of 65 outcomes with a conclusion of an effect (statistically significant difference), 18 (28%; 95% CI 18–40%) had an increased risk for a type I error.

Outcomes measuring harms had increased risks for random errors more frequently than outcomes measuring benefits (Fig. 2). Overall, 47% (95% CI 35–58%) of outcomes assessing harms (k = 75) were affected, compared with 12% (95% CI 4–32%) of outcomes reporting benefits (k = 25).

We conducted two sensitivity analyses to explore the robustness of our results (Table 3). We raised the type II error to 20% and used effect measures and statistical models exactly as used in Cochrane reports. Overall, proportions of outcomes with increased risks for random errors did not change substantially.

3.2. Exploring reasons for increased risks for type I or type II errors

3.2.1. Lack of adherence to guidance

Of 38 outcomes with increased risks for random errors according to TSA, the reassessment of the CoE showed that 28 outcomes (74%) should not have been graded as high CoE based on current GRADE guidance. Most commonly, Cochrane reviewers deviated from guidance when grading imprecision (24 cases).

For example, in one review, investigators examined the comparative benefits and harms of transmyocardial laser revascularization vs. medical therapy for refractory angina [20]. Based on a meta-analysis of seven RCTs with 1,053 participants, authors concluded with high CoE that both treatment options have similar risks of mortality at 1 year (odds ratio 1.12; 95% CI 0.77–1.63). Although the effect estimate does not show a statistically significant difference in mortality, the upper limit of the CI encompasses a difference in the risk for mortality that most decision-makers would probably view as clinically relevant (i.e., a 63% relative increase of mortality). Furthermore, the number of deaths (n = 123) across all trials in this meta-analysis was below the required information size. The required information size, given a baseline risk of death of 11.9% (calculated from the control group) and a relative risk difference of 25%, would be about 5,000 participants. In such a case, current GRADE guidance recommends downgrading for imprecision [8].

An extreme case of a mistaken grade of high CoE was a review on postexposure passive immunization with gamma globulin vs. control for preventing rubella [21]. Authors graded the outcome of "number of patients with rubella at 6–8 months" as high CoE although the available three studies included only 37 participants (21 developing rubella).

Table 3. Different types of analyses and corresponding proportions of outcomes with increased random error

Analyses	Outcomes with increased risk for random errors (%)	Comments and interpretation
Per protocol analysis	38%	Analysis as specified in the study's protocol and reported in Methods with relative risk as outcome measure and DerSimonian and Laird's random effects model when appropriate; $\alpha = 5\%$; $\beta = 10\%$; fixed effect model in cases in which random effects were not appropriate.
Changing threshold for type II error to 20%	33%	Some analysts view a type II error of 20% as acceptable. We used the per protocol analysis except that we applied a less conservative type II error of 20%. In five cases, the risk for random errors was not increased anymore.
Using the same effect measures and statistical models as used in Cochrane reviews	35%	The choice of the outcome measure (odds ratio or relative risk) or the choice of the statistical model (random or fixed effects) can affect statistical significance or type I or type II errors. We used exactly the same outcome measures and statistical models that Cochrane authors used when they graded the CoE. In three cases, the risk for random errors was not increased anymore.

3.2.2. Predictors of increased risks

To explore factors that might predict increased risks for type I or type II errors, we conducted multivariate regression analyses.

We detected two statistically significant predictors of increased risks for random errors: small absolute risk difference ($P = 0.009$) and few events in a meta-analysis ($P = 0.001$). Type of outcome (harmful or beneficial), number of studies in a meta-analysis, type of intervention, baseline risk, and statistical significance of results were not significantly associated with an increased risk for random errors. A large magnitude of treatment effect (defined by GRADE as a relative risk of >2 or <0.5) significantly predicted that the risk for random errors would not be increased ($P = 0.012$).

Fig. 3 depicts the risk for increased random errors by absolute risk differences based on an extrapolation of results from the multivariate regression analysis. Outcomes with small absolute risk differences ($<10\%$ points) had a more than 70% risk of having an increased risk for random error. The risk fell below 5% when the absolute risk difference was larger than 27 percentage points.

4. Discussion

4.1. Overview of findings

To our knowledge, our work is the first attempt to assess whether outcomes rated as high CoE in systematic reviews have increased risks for random errors. Our findings revealed that 38% of high CoE outcomes in randomly selected systematic reviews, which all happened to be Cochrane reviews, had higher risks for type I or type II errors than is conventionally acceptable. These results are particularly worrisome because decision-makers who rely on systematic reviews expect that outcomes rated "high CoE" provide a solid evidence base for decisions. Decision-makers rely on these assessments to develop clinical practice guidelines or other recommendations, make clinical or health policy decisions, or create decision aids or patient information materials. Outcomes rated as high CoE are often the justification for strong recommendations in clinical guidelines and for insurance coverage, public health, and other important health policy decisions. Finding that more than one-third of outcomes graded as high CoE have an increased risk for being false-positive or false-negative is concerning.

Exploration of the reasons for the increased risk of random errors revealed that in a considerable majority of cases (78%), investigators rating the CoE did not adhere to current guidance, particularly to guidance about imprecision. Most likely, many of these outcomes should not have received a high CoE grade. Grades of CoE always have an element of subjective judgment, guidance for the imprecision domain, however, offers a framework for downgrading that is sometimes challenging to implement but relies on

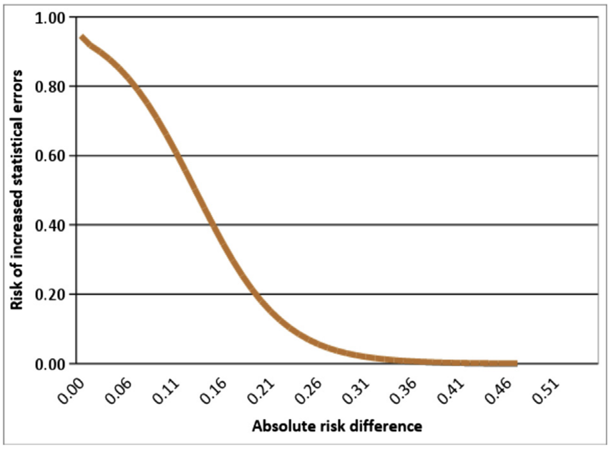


Fig. 3. Absolute risk difference as a predictor of random error. Likewise, the number of events in a meta-analysis (Table 3) was strongly associated with an increased risk for random error. Outcomes with fewer than 700 events that had been graded as high CoE had a more than 45% probability of having an increased risk for random errors (Fig. 4).

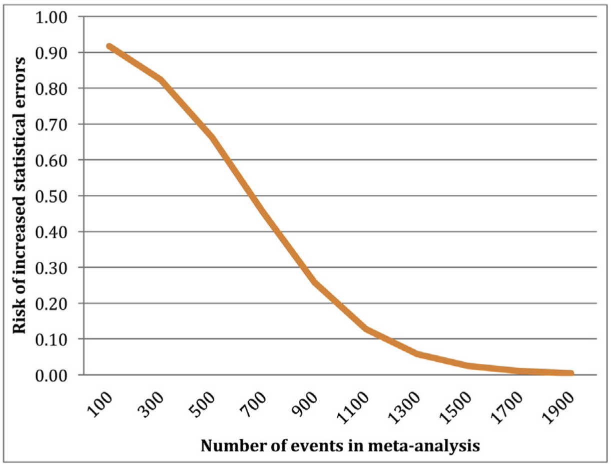


Fig. 4. Number of events in meta-analyses as a predictor of random error.

quantitative thresholds that can be calculated [8]. Better guidance, training, and quality assurance during review production might have prevented some of the incorrect grades.

In 22% of the 38 discordant outcomes, reviewers adhered to current guidance; nevertheless, outcomes still had increased risks for type I or type II errors. This finding implies that the conceptual approach to grading CoE has deficiencies. The current approach may not adequately take into consideration between-study heterogeneity and repeated updates of meta-analyses. Methods research shows that these factors can raise the risks for type I and type II errors [5,7]. In particular, based on our results, guidance might not be adequate for studies with small absolute risk differences or with fewer than 700 events in a meta-analysis. This observation is in accordance with simulation studies by Thorlund et al. [22]. Frequently, authors of systematic reviews drew a conclusion of no effect based on a nonsignificant result with an increased risk of being false-negative (type II error). This was particularly common for outcomes assessing harms. In some cases, results will indeed be false-negative; when this is the case, decision-makers using the systematic reviews might draw incorrect inferences about the balance of benefits and harms.

The fact that 75% of randomly selected outcomes rated as high CoE assessed harms might indicate that systematic review authors are more careless in assigning high CoE grades when drawing conclusions about harms, particularly when drawing conclusions about similar risks of harmful effects. Negligence of harms is common in clinical research [23,24].

Our findings are consistent with other methods research assessing random errors in meta-analyses. Turner et al. found that most Cochrane meta-analyses were underpowered [25]. In a series of cumulative meta-analyses, Imberger et al. detected false-positive findings in 7% of Cochrane meta-analyses [26]. Applying TSA to systematic reviews of anesthesiologic interventions revealed that only 12% of meta-analyses had a power of 80% or higher [27]. Likewise, 72% of apparently conclusive meta-analyses on neonatal topics had increased risks of random errors [18]. A recent review of 100 Cochrane systematic reviews revealed low adherence of authors to GRADE guidance when grading imprecision. In addition, most authors did not adequately report reasons for downgrading for imprecision [28]. None of these studies, however, focused on high CoE outcomes that, by definition, should not have increased risks of random errors because they are the cornerstone of clinical and health policy decision-making.

4.2. Potential study limitations

First, our sample might not be representative of most systematic reviews that grade CoE. We randomly sampled Cochrane and AHRQ reviews because both institutions are known for rigorous methods and extensive internal and external peer reviews [29,30]. Conceivably, the lack of adherence to guidance documents might be more prevalent in systematic reviews that do not use such high methodological standards.

Second, TSA modeling requires choices in effect sizes that are the basis for the calculation of the required information size. We chose general thresholds for clinical relevance based on our clinical understanding of outcomes; still, being certain that these choices are correct is impossible. For example, we used a relative risk reduction of 10% for outcomes that we deemed highly patient-relevant (e.g., mortality, pain, serious adverse events) and 25% for other outcomes that may be less important to patients. Clinicians or patients might reasonably disagree with these choices.

Third, we used DerSimonian and Laird's random effects models for most of our TSAs (in one case, we used a fixed effect models because the random effects model did not seem appropriate), although TSA offers other options. We chose the DerSimonian and Laird's [31] method because it is still the default choice in most software packages. We are aware of the limitations of this method for meta-analyses of inconsistent effects [32].

Fourth, we used relative risks as the measure of association for all meta-analyses even if authors of an included review used odds ratios. In sensitivity analyses, we explored the impact of the choice of the statistical model and the measure of association. In a few cases, results regarding increased risks of random error changed, but the overall finding remained the same.

Finally, TSA focuses only on random error. Systematic error (bias) and selective reporting of outcomes and studies can have a substantial impact on effect estimates. In our study, we relied on risk of bias and selective reporting assessments that the Cochrane authors made. Because author groups differed across these systematic reviews, heterogeneity in approaches and varying adherence to the Cochrane Handbook regarding risk of bias is likely [11].

4.3. Approaches to assessing certainty of estimates

Over the past decade, GRADE and closely related systems such as that of the EPC program of AHRQ have evolved as widely used approaches to convey the certainties and uncertainties inherent in research. Their conceptual frameworks use information about factors that most researchers would intuitively consider when assessing the confidence in findings based on a body of evidence. Compared with other approaches, these systems have clear advantages because they make decisions about the CoE transparent and explicit [33].

An increased risk for random errors in more than one-third of outcomes graded as high CoE is, however, concerning. Partially, this conclusion is grounded in the concept of GRADE but mostly in the way the instrument is operationalized. The GRADE Working Group needs to reflect on how to reduce unwarranted variation in the use of GRADE, but it also may want to reconsider its current approach to rating imprecision. Jakobsen et al. suggest how results of TSA could be implemented in rating CoE [34]. In addition, incorporating a tool such as TSA into the GRADEpro Guideline Development Tool (www.guidelinedevelopment.org) would give investigators valuable information about the required information size. This step might lead to more consistent assessments, especially of the imprecision domain.

Finally, future methods research needs to explore how systematic reviewers can better capture uncertainty in meta-analyses that produce only small absolute risk reductions. Of particular importance would be exploring these factors in the context of current approaches to grading CoE.

Acknowledgments

The authors gratefully acknowledge the Cochrane Editorial Unit for providing data files of the Cochrane reviews. The authors deeply appreciate the considerable support efforts from the following individuals: RTI International Staff (Kathleen N. Lohr, PhD, Carol Woodell, BSPH, and Loraine Monroe); Danube University Krems (Petra Grob, MBA, Irma Klerings, and Sandra Hummel). Finally, they thank Kylie Thaler, MD, MPH, for help with grading.

This research was funded under grant number 1RO3HS024749-01 from the Agency for Healthcare Research and Quality (AHRQ), U.S. Department of Health and Human Services, United States.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2018.10.009.

References

[1] Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64:383–94.

[2] Berkman ND, Lohr KN, Ansari M, McDonagh M, Balk E, Whitlock E, et al. Grading the strength of a body of evidence when assessing health care interventions for the effective health care program of the Agency for Healthcare Research and Quality: an update. Methods Guide for Comparative Effectiveness Reviews (Prepared by the RTI-UNC Evidence-based Practice Center under Contract No. 290- 2007-10056-I). AHRQ Publication No. 13(14)-EHC130-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2013. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

[3] Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. The GRADE Working Group clarifies the construct of certainty of evidence. J Clin Epidemiol 2017;87:4–13.

[4] Gartlehner G, Dobrescu A, Evans TS, Bann C, Robinson KA, Reston J, et al. The predictive validity of quality of evidence grades for the stability of effect estimates was low: a meta-epidemiological study. J Clin Epidemiol 2016;70:52–60.

[5] Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. J Clin Epidemiol 2008;61:763–9.

[6] Thorlund K, Devereaux PJ, Wetterslev J, Guyatt G, Ioannidis JP, Thabane L, et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? Int J Epidemiol 2009;38:276–86.

[7] Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size by quantifying diversity in random-effects model meta-analyses. BMC Med Res Methodol 2009;9:86.

[8] Guyatt G, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence-imprecision. J Clin Epidemiol 2011;64:1283–93.

[9] Borm GF, Donders AR. Updating meta-analyses leads to larger type I errors than publication bias. J Clin Epidemiol 2009;62:825–830 e10.

[10] Hu M, Cappelleri JC, Lan KK. Applying the law of iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes. Clin Trials 2007;4:329–40.

[11] Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]. The Cochrane Collaboration; 2011. Available from: www.handbook.cochrane.org. Accessed February 11, 2018.

[12] Thorlund K, Engstrom J, Wetterslev J, Brok J, Imberger G, Gluud C. Trial sequential analysis (TSA): user manual. Copenhagen: Centre for Clinical Intervention Research; 2011.

[13] Rothman K, Greenland S, Lash T. Modern epidemiology. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilki, 3rd revised edition; 2013.

[14] Wetterslev J, Jakobsen JC, Gluud C. Trial sequential analysis in systematic reviews with meta-analysis. BMC Med Res Methodol 2017;17:39.

[15] O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35:549–56.

[16] Gordon Lan KK, Demets DL. Discrete sequential boundaries for clinical trials. Biometrika 1983;70:659–63.

[17] Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;64:401–6.

[18] Brok J, Thorlund K, Wetterslev J, Gluud C. Apparently conclusive meta-analyses may be inconclusive–Trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses. Int J Epidemiol 2009;38:287–98.

[19] Sun R, Jia WQ, Zhang P, Yang K, Tian JH, Ma B, et al. Nitrous oxide-based techniques versus nitrous oxide-free techniques for general anaesthesia. Cochrane Database Syst Rev 2015;11:CD008984.

[20] Briones E, Lacalle JR, Marin-Leon I, Rueda JR. Transmyocardial laser revascularization versus medical therapy for refractory angina. Cochrane Database Syst Rev 2015;2:CD003712.

[21] Young MK, Cripps AW, Nimmo GR, van Driel ML. Post-exposure passive immunisation for preventing rubella and congenital rubella syndrome. Cochrane Database Syst Rev 2015;9:CD010586.

[22] Thorlund K, Imberger G, Walsh M, Chu R, Gluud C, Wetterslev J, et al. The number of patients and events required to limit the risk of overestimation of intervention effects in meta-analysis–a simulation study. PLoS One 2011;6:e25491.

[23] Ioannidis JP. Adverse events in randomized trials: neglected, restricted, distorted, and silenced. Arch Intern Med 2009;169:1737−9.

[24] Storebo OJ, Pedersen N, Ramstad E, Kielsholm ML, Nielsen SS, Krogh HB, et al. Methylphenidate for attention deficit hyperactivity disorder (ADHD) in children and adolescents - assessment of adverse events in non-randomised studies. Cochrane Database Syst Rev 2018; 5:CD012069.

[25] Turner RM, Bird SM, Higgins JPT. The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. PLoS One 2013;8:e59202.

[26] Imberger G, Thorlund K, Gluud C, Wetterslev J. False-positive findings in Cochrane meta-analyses with and without application of trial sequential analysis: an empirical review. BMJ Open 2016;6:e011890.

[27] Imberger G, Gluud C, Boylan J, Wetterslev J. Systematic reviews of anesthesiologic interventions reported as statistically significant: problems with power, precision, and type 1 error protection. Anesth Analg 2015;121:1611−22.

[28] Castellini G, Bruschettini M, Gianola S, Gluud C, Moja L. Assessing imprecision in Cochrane systematic reviews: a comparison of GRADE and trial sequential analysis. Syst Rev 2018;7(1):110.

[29] Higgins JPT, Lasserson T, Chandler J, Tovey D, Churchill R. Methodological Expectations of Cochrane Intervention Reviews (MECIR) Standards for the conduct and reporting of new Cochrane Intervention Reviews, reporting of protocols and the planning, conduct and reporting of updates Version 1.0. 2 ed. London: Cochrance Methods; 2016.

[30] Jorgensen AW, Hilden J, Gotzsche PC. Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs: systematic review. BMJ 2006;333:782.

[31] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177−88.

[32] Cornell JE, Mulrow CD, Localio R, Stack CB, Meibohm AR, Guallar E, et al. Random-effects meta-analysis of inconsistent effects: a time for change. Ann Intern Med 2014;160:267−70.

[33] Schunemann HJ, Best D, Vist G, Oxman AD. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. CMAJ 2003;169:677−80.

[34] Jakobsen JC, Wetterslev J, Winkel P, Lange T, Gluud C. Thresholds for statistical and clinical significance in systematic reviews with meta-analytic methods. BMC Med Res Methodol 2014;14:120.