Journal of Clinical Epidemiology 114 (2019) 1−10

Journal of Clinical Epidemiology

REVIEW

Apparently conclusive meta-analyses on interventions in critical care may be inconclusive—a meta-epidemiological study

Thijs M. Koster[a,*], Jørn Wetterslev[b], Christian Gluud[b], Janus C. Jakobsen[b,c], Thomas Kaufmann[d], Ruben J. Eck[e], Geert Koster[a], Bart Hiemstra[d], Iwan C.C. van der Horst[a], Eric Keus[a]

[a]*Department of Critical Care, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands*
[b]*Department 7812, Rigshospitalet, The Copenhagen Trial Unit (CTU), Centre for Clinical Intervention Research, Copenhagen University Hospital, DK-2100 Copenhagen, Denmark*
[c]*Department of Cardiology, Holbæk Hospital, Holbæk, Denmark*
[d]*Department of Anesthesiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands*
[e]*Department of Internal Medicine, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands*

Accepted 18 May 2019; Published online 11 June 2019

Abstract

Objectives: Risks of random type I and II errors are associated with false positive and false negative findings. In conventional meta-analyses, the risks of random errors are insufficiently evaluated. Many meta-analyses, which appear conclusive, might, in fact, be inconclusive because of risks of random errors. We hypothesize that, for interventions in critical care, false positive and false negative findings frequently become apparent when accounting for the risks of random error. We aim to investigate to which extent apparently conclusive conventional meta-analyses remain conclusive when adjusting statistical significance levels and confidence intervals considering sparse data and repeated testing through Trial Sequential Analysis (TSA).

Methods: We searched The Cochrane Library, MEDLINE, and EMBASE for reviews of interventions in critical care. We used TSA with the relative risk reduction from the estimated meta-analyzed intervention effects adjusted for heterogeneity based on the observed diversity. We report proportions of meta-analyses and potential inconclusive findings of positive, neutral, and negative conclusions based on conventional naïve meta-analyses, which use an alpha of 5% and 95% confidence intervals. In TSA-controlled meta-analyses showing a beneficial or harmful intervention effect, we assessed the risk of bias by six Cochrane domains.

Results: A total of 464 reviews containing 1,080 meta-analyses of (co-)primary outcomes were analyzed. From the 266 conventional meta-analyses suggesting a beneficial effect, 133 (50%) were true positive and 133 (50%) were potentially false positive according to TSA. From the 755 conventional meta-analyses suggesting a neutral effect, there were 214 (28%) true neutral and 541 (72%) were potentially false neutral according to TSA. From the 59 conventional meta-analyses suggesting a harmful effect, 17 (29%) were true negative and 42 (71%) were potentially false negative according to TSA. When the true beneficial and true harmful meta-analyses according to TSA were evaluated for risk of bias, new TSAs conducted on only trials with overall low risk of bias showed only firm evidence of a beneficial effect on one outcome and a harmful effect on one outcome.

Conclusions: Of all meta-analyses in critical care, a large proportion may reach false conclusions because of unknown risks of random type I or type II errors. Future critical care meta-analyses should aim for establishing an effect of interventions accounting for risks of bias and random errors.

Keywords: Meta-epidemiological study; Trial sequential analysis; Meta-analysis

1. Introduction

Systematic reviews with meta-analyses of randomized clinical trials (RCTs) are considered the highest level of evidence for intervention research [1,2]. However, systematic reviews might lose credibility when analyses and conclusions are invalid because of the risks of errors either in the systematic review process or included RCTs [1,3]. Evidence to

Funding: No funding was received.
Conflicts of interests: All authors declare no conflicts of interests.
Patients' contribution: Patients were not involved in the development or conduct of the study.

* Corresponding author. Tel.: +31 50 361 6161; fax: +31 50 361 5644.
E-mail address: t.m.koster@umcg.nl (T.M. Koster).

https://doi.org/10.1016/j.jclinepi.2019.05.011
0895-4356/

What is new?

Key findings

- At least 50% of all meta-analyses on interventions in critical care, which claim a significant intervention effect based on conventional statistical methods, are potentially false positive when accounting for random error using Trial Sequential Analysis.
- Also, 80% of all meta-analyses on interventions, which claim a nonsignificant intervention effect, may be false neutral.

What this adds to what was known?

- Interventions in daily practice of critical care may either be falsely accepted or false rejected due to insufficient evidence.

What is the implication and what should change now?

- Future studies in critical care should increase their efforts to address the risks of random errors.

support interventions in critical care patients is limited and subjected to risks of bias and risk of random errors [4,5]. In a previous article, we observed that according to the Risk Of Bias In Systematic reviews tool, only 0.9% of all meta-analyses on interventions in critical care were conducted within a systematic review process, which appeared to have fulfilled the criteria associated with a systematic review conducted with low risk of bias [6,7]. Random errors (or ''the play of chance'') of the accumulated data in meta-analyses may be another major reason for misleading results [8−10]. A fixed-effect model meta-analysis should include an information size at least as large as the sample size of an adequately powered single RCT to reduce the risks of type I and type II random errors [11,12]. In addition, the required information size in a meta-analysis (the ''meta-analytic sample size'') should be adjusted for statistical heterogeneity as more information is required when heterogeneity increases [9,10,13]. Trial Sequential Analysis (TSA) combines an estimated required information size for a meta-analysis with the adaptation of monitoring boundaries to evaluate accumulating meta-analytic updates [9,10,12]. In this context, TSA may serve as a tool for quantifying the reliability of cumulative data in meta-analyses [9,10,12].

We hypothesize that many of the conclusive conventional meta-analyses on interventions in critical care will appear inconclusive when accounting for potential risks of random errors and risks of bias. The aim of this study was to evaluate the risks of random errors in all meta-analyses of interventions in critical care and explore the extent to which conclusive traditional meta-analyses remain conclusive when accounting for potential risk of random error because of sparse data and repetitive testing through TSA. We will also evaluate how many of the statistically insignificant meta-analyses actually have the power to exclude an anticipated realistic intervention effect.

2. Methods

This meta-epidemiologic study was conducted following our prepublished protocol and the addendum (Supplemental data, Appendices A and B).

Throughout this report, we use the term review to refer to either systematic or nonsystematic reviews with meta-analytic assessments, which were all included in this study, whereas many may not qualify for the highest level of evidence [14,15].

2.1. Selection criteria

We considered all reviews eligible for inclusion, irrespective of language, quality of conduct, or quality of reporting, suggested respectively by The Cochrane Handbook for Systematic Reviews of Interventions and the PRISMA statement [6,14,16]. We used no restrictions on the numbers of meta-analyses for each type of intervention and used only the most recent version in case of updated versions (i.e., Cochrane reviews). We excluded reviews of observational studies and observational data meta-analyzed along with data from RCTs as well as indirect comparisons in network meta-analyses [16].

We selected reviews including RCTs on critically ill adult patients (aged $\geq$18 years). Reviews that included RCTs of both adults and children were included, but only data on adults were extracted. Critical illness encompassed any clinical setting wherein patients required treatment at the intensive care unit (ICU).

Only reviews that evaluated interventions used for patients in critical care and either performed or authorized by intensivists were included. For example, inotropic agents or vasopressors are typically initiated by intensivists. Interventions primarily initiated by other medical specialists were excluded if discontinued at the ICU. For example, a decompressive craniectomy is conducted by the neurosurgeon, yet these patients may be admitted to the ICU thereafter. Interventions initiated in another setting but continued in the ICU were included, for example, early goal-directed therapy or antibiotic therapy initiated in the emergency room and continued in the ICU. Some interventions are most frequently used in wards and incidentally also in the ICU; these interventions were excluded when the majority was not applied in the ICU, for example, early enteral feeding in patients suffering pancreatitis.

We included all reviews independent of the control group intervention, hence those including an active comparator, placebo, usual care, or no intervention.

We focused on outcomes critical or important for decision-making according to the perspective of patients as recommended by the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group [16−18]. Patient-important outcomes are typically dichotomous (e.g., mortality) and less frequent subjective. To enhance objectivity of the outcomes and guarantee the patient centeredness, we included all meta-analyses that reported at least one dichotomous outcome.

2.2. Search strategy and data extraction

We searched the databases *"Cochrane Reviews"* and "*Other Reviews"* in *The Cochrane Library*, *MEDLINE*, and *EMBASE* using sensitive search strategies (Supplemental data, Appendix C). Selections were performed by two investigators; any uncertainties were resolved through discussion. For each review, we recorded author, year of publication, numbers of included RCTs and randomized patients, the intervention evaluated, and the (co-)primary dichotomous outcomes up to a maximum of three.

2.3. Intuitive explanation of TSA

Most meta-analyses in systematic reviews do not have sufficient statistical power to detect or refute even large intervention effects. This is why a meta-analysis ought to be regarded as an interim analysis on its way toward the required (sufficient) information size [19]. TSA offers adjusted confidence intervals (CIs) and restricted thresholds for statistical significance when the diversity-adjusted required information size for the meta-analysis has not been reached. As long as the cumulative z-statistic has not crossed any of the adjusted monitoring boundaries for benefit, harm, or futility, there is still a higher risk (than, e.g., 5% type I error or 20% type II error) that with more accumulating data conclusions will change: early significant findings may turn insignificant, and reverse, early insignificant effects may become significant (either beneficial or harmful). In short, TSA provides a frequentistic approach to control both type I and type II errors considering sparse data and repetitive testing. TSA is therefore a method analog to Group Sequential Analysis of single trials with interim analyses (Group Sequential Design) on their way to include a preplanned sample size. The TSA confidence limits are therefore not set at 95%, but adjusted (always higher than 95% unless the RIS has been reached) according to the acquired information size relative to the RIS and the corresponding trial sequential monitoring boundaries, LanDemets group sequential boundaries [10,19−21].

2.4. Data analysis

All data were analyzed using the Copenhagen Trial Unit's computer program, TSA version 0.9.5.10 beta (www.ctu.dk/tsa).

The TSA v0.9 graphically displays the relationship between the cumulative *z*-score, the information size, and the two-sided trial sequential monitoring boundaries analog to the Lan-DeMets group sequential boundaries provided that the accrued number of participants constitute more than 4% of the required number of participants (Fig. 1) [11]. We calculated both the conventional 95% CI and the TSA-adjusted CI with corresponding *P* values with the DerSimonian-Laird random effects model [9,22,23]. We defined that there was firm evidence for an intervention effect when the cumulative *z*-curve crossed one of the monitoring boundaries for benefit, futility, or harm.

TSA may also indicate that an intervention is unlikely to have the anticipated effect. Futility boundaries (or inner wedge) are a set of thresholds that reflect the uncertainty of obtaining a chance negative finding in relation to the strength of the available evidence (Fig. 1) [10,24]. Within the threshold, the test statistic (z-value) is so low that the likelihood of a significant beneficial or harmful effect being found becomes negligible. However, one might claim a smaller intervention effect, if clinically relevant, and then more RCTs and patients may be needed, that is, the required information size increases, and the futility boundaries shift to the right accordingly.

We used relative risk, and the control event rate was estimated from the summarized unweighted events in the control group. We conducted sensitivity analyses using odds ratios in case the event proportion was below 5% in the control group. For single zero event trials (zero events in only one of the compared groups), we made empirical adjustments of 0.01 [25]. Heterogeneity increases the uncertainty in meta-analyses and is commonly measured using the inconsistency factor I^2 [26]. In any case where the trial weights are not equal, using I^2 will lead to an underestimation of the adjustment factor, and thus, an underestimation of the required information size. Alternatively, heterogeneity may also be estimated by diversity (D^2), which seems a better alternative than I^2 to consider model variation in any random-effects meta-analysis [20,27]. D^2 is especially constructed to account for the statistical heterogeneity in the calculation of the required information size (given an a priori anticipated intervention effect), and this is not the case with I^2, which does not account for all the variation increase going from a fixed-effect model to a random-effects model [20]. Therefore, TSA is programmed to use D^2, and for each meta-analysis, we calculated the diversity-adjusted required information size [9]. The diversity used in the analyses was either as estimated in the meta-analysis or an anticipated value.

For each meta-analysis, we conducted the conventional analysis, one primary TSA, and three additional sensitivity TSAs to test the robustness of our findings.

2.4.1. Primary analyses

The primary TSA was conducted based on the observed relative risk reduction (RRR) and the observed D^2 in the

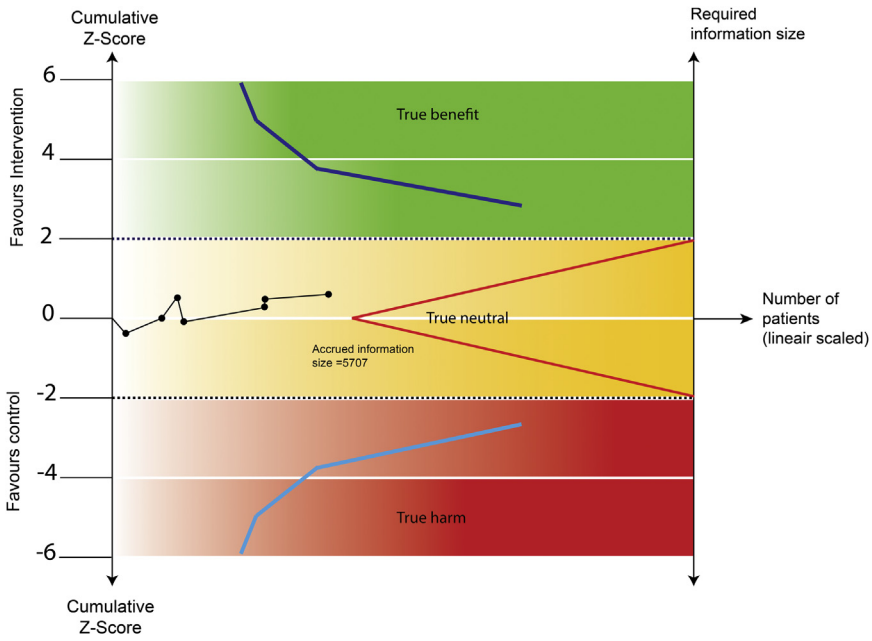

Fig. 1. Example of Trial Sequential Analysis showing absence of evidence for an intervention effect on the meta-analyzed outcome. The black dotted lines represent the conventional boundaries for benefit and harm ($P < 0.05$). The dark blue line represents the trial sequential boundary for benefit. The light blue line represents the trial sequential boundary for harm. The red lines represent the futility boundaries. The black line represents the cumulative z-curve of the meta-analyzed outcome. The black dots on this line represent the additional of another trial to the cumulative data of this outcome. If the cumulative z-curve crosses the trial sequential boundary for benefit, it can be concluded that the intervention has a true benefit effect, and accordingly, crossing the trial sequential boundary for harm suggests a true harmful intervention effect. If the cumulative z-curve crosses one of the futility boundaries, it can be concluded that the intervention does not possess the anticipated intervention effect compared with the control intervention. In this example, the cumulative Z-curve does not cross the conventional boundaries, neither the trial sequential boundaries nor futility boundaries; thus, there is insufficient data to accept or reject the anticipated intervention effect. In this example, the intervention effect may therefore still be potentially false neutral (the light yellow area). Concordantly, the light green area and light red area represent areas where the cumulative z-values is potentially false positive and potentially false negative. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

meta-analysis with an overall type I error (α) of 0.05 and a type II error (β) of 0.20 (power 80%).

2.4.2. Sensitivity analyses

We performed a first sensitivity TSA with an a priori *anticipated* D^2 of 25% as our best guess of a heterogeneity adjustment when the meta-analysis eventually reaches its required information size and the actual observed RRR estimated for each outcome [20,28].

We conducted a second sensitivity TSA using an a priori *anticipated* RRR of 25% and the actual observed D^2 because early testing with sparse data may overestimate the intervention effect. On the other hand, the ''true'' intervention effect may eventually appear to be higher (e.g., around 25% RRR) than an initially underestimated intervention effect (e.g., 10%).

The third sensitivity TSA was conducted with an RRR of 25% and a D^2 of 25%.

We performed these sensitivity analyses because of the uncertainty regarding estimations of anticipated intervention effects and heterogeneity. If a meta-analysis did not cross the monitoring boundaries in both the primary and the sensitivity analyses, the meta-analysis was considered to be potentially false positive, potentially false negative, or potentially false neutral as the results were inconclusive.

2.5. Assessment of bias risk

The TSA can conclude that there is a ''truly'' significant intervention effect if the trial sequential monitoring boundary for benefit or harm has been crossed when all included trials in the meta-analysis are at overall low risk of bias. Because any risk of bias would challenge a significant result for benefit or harm, we assessed the risks of bias of the RCTs included in the meta-analyses crossing the trial sequential monitoring boundary for benefit or harm, using the Cochrane tool for bias risk assessment for the domains of allocation sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessors, incomplete outcome data, and selective outcome reporting [6]. We classified trials at overall low risk of bias if all the domains were judged low risk of bias. Trials with one or more domains judged at unclear or high risk of bias were considered to be at overall high risk of bias.

For meta-analyses that did not cross the trial sequential monitoring boundary for benefit or harm (including those meta-analyses which crossed the futility boundary), we

concluded that there was insufficient evidence to recommend or reject the intervention, and we did not assess the risk of bias of the trials in these meta-analyses.

We performed subgroup analyses according to bias risk assessment only in meta-analyses that crossed the trial sequential monitoring boundary for benefit or harm in all the primary and sensitivity analyses.

3. Results

The search strategy identified 23,882 hits, of which 713 remained after screening (Fig. 2). Full-text evaluation excluded another 224 (Supplemental Data, Table 1), whereas 489 hits fulfilled our inclusion criteria. Because limited translation resources, we had to exclude 25 hits (Chinese $n = 23$; Spanish $n = 2$). Accordingly, 464 reviews were included in this meta-epidemiologic study (a list of included reviews is attached as E-component).

The included 464 reviews evaluated a wide variety of interventions (Supplemental data, Table 2) with an overall total of 1,080 (co-)primary outcomes (three, two, and one (co-)primary outcomes were reported by 253, 110, and 101 meta-analyses, respectively). For each meta-analyzed outcome, data were provided by a median of six RCTs (interquartile range [IQR] 3−10) with a median total of 966 participants (IQR 436−1,907).

A statistically significant beneficial intervention effect with a $P < 0.05$ threshold was suggested in 266 (25%) of 1,080 meta-analyses using conventional methods and a statistically significant harmful intervention effect in 59 (5.5%) meta-analyses (Fig. 3). A total of 755 (70%) out of 1,080 meta-analyses suggested a neutral effect.

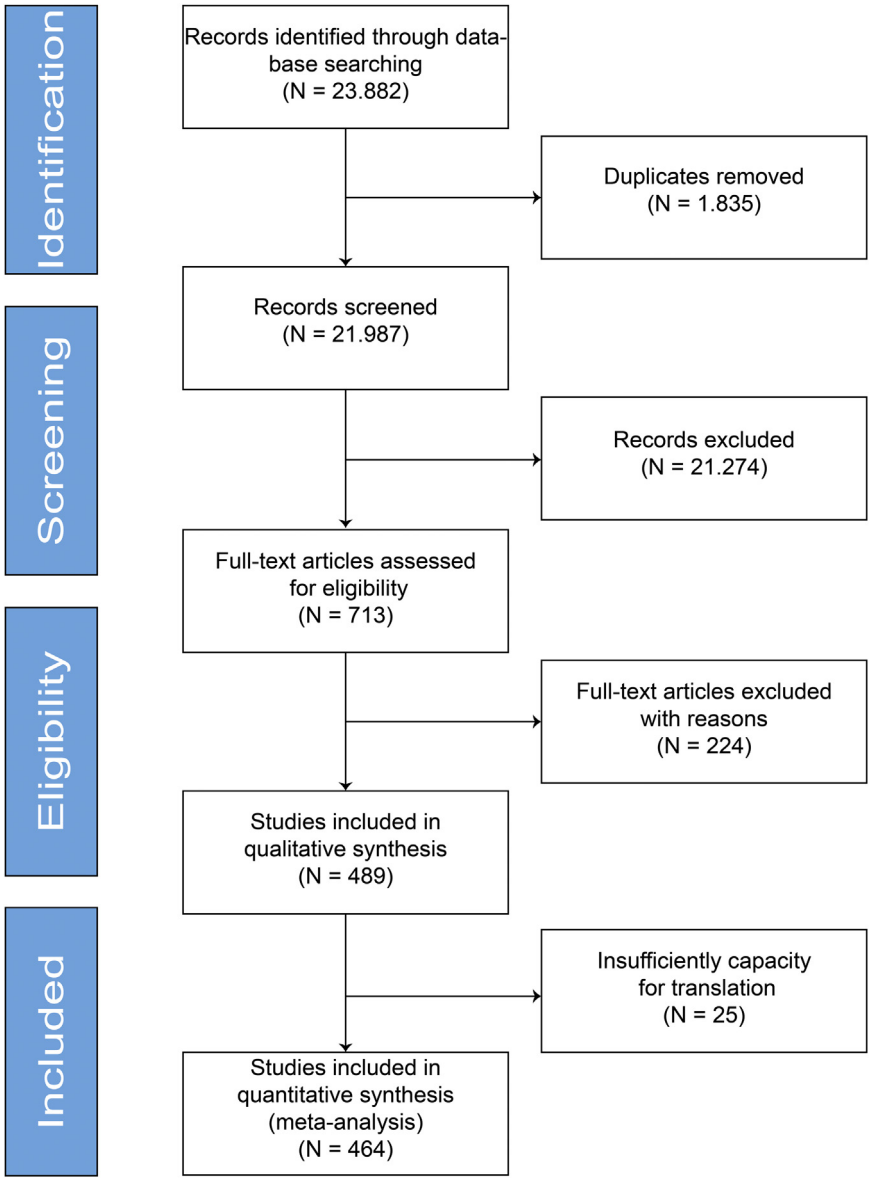

Fig. 2. Flow chart of study selection according to PRISMA. Reasons for exclusion based on full text are given in Supplemental Table 2.

3.1. Primary TSA: RRR estimated and D2 estimated in the meta-analysis

From the 266 meta-analyses that suggested a beneficial intervention effect based on an alpha of 5%, 133 (50%) were true positive and 133 (50%) potentially false positive according to TSA (Fig. 3, Table 1). From the 755 meta-analyses that suggested a neutral intervention effect based on an alpha of 5%, there were 214 (28%) true neutral and 541 (72%) potentially false neutral according to TSA. From the 59 meta-analyses that suggested a harmful intervention effect based on an alpha of 5%, there were 17 (29%) true negative and 42 (71%) potentially false negative according to TSA. In total, out of 1,080 conventional meta-analyses, TSA suggested a premature conclusion in 716 (66%).

3.2. Sensitivity analyses

The proportions of true positive meta-analyses ranged from 10.4% to 12.8% and potentially false positive ranged from 12.2% to 14.8% (Table 1). The proportions of true neutral meta-analyses ranged from 19.1% to 23.7% and potentially false neutral range from 45.4% to 50.7%. The proportions of true negative meta-analyses ranged from 1.4% to 1.9% and potentially false negative ranged from 3.2% to 4.3%. The total potentially false conclusions suggested by TSA ranged from 64% to 66%.

In 70 meta-analyses, control event proportions were <5%. TSA sensitivity analyses using odds ratio instead of RRR (with D^2 estimated) changed conclusions in 22 (31%) out of 70 meta-analyses.

3.3. Conclusions based on all four TSA scenarios

TSA assessed 43 (4.0%) meta-analyses as true positive in all four scenarios (one primary and three sensitivity analyses) and 216 (20%) in at least one scenario. TSA assessed 150 (14%) meta-analyses as true neutral in all four scenarios and 216 (20%) in at least one scenario. TSA assessed one (0.01%) meta-analysis as true negative in all four scenarios and 43 (4.0%) in at least one scenario.

Of all conventional meta-analyzed statistically significant outcomes (either beneficial or harmful intervention effects), 281 of 325 (87%) meta-analyses appeared inconclusive (in the primary and sensitivity TSA). Moreover, 605 of 755 (80%) neutral meta-analyses appeared inconclusive. Thus, 886 of 1,080 (82%) meta-analyses appeared inconclusive.

3.4. Complementary analysis of risk of bias of RCTs included in meta-analyses with beneficial or harmful significant intervention effects according to all TSAs

A total of 44 meta-analyses (including a total of 438 individual RCTs) crossed the trial sequential monitoring boundary for benefit ($n = 43$) or harm ($n = 1$) in both the primary *and* all three sensitivity TSA. The full text of eight articles could not be retrieved. Fourteen articles

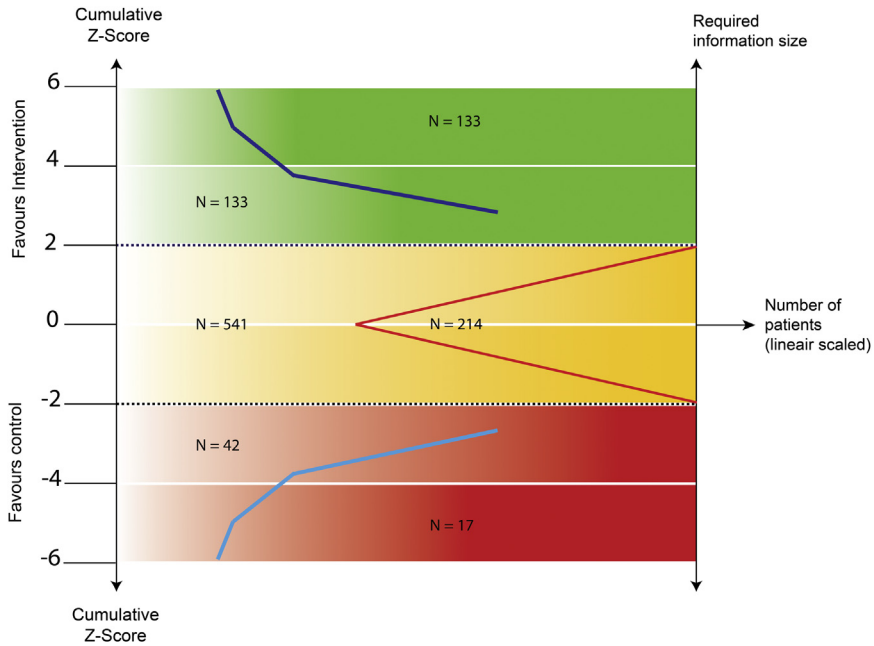

Fig. 3. Results of the primary analysis. The dark blue line represents the trial sequential boundary for benefit. The light blue line represents the trial sequential boundary for harm. The red lines represent the futility boundaries. The black line represents the cumulative z-curve of the meta-analyzed outcome. The black dots on this line represent the additional of another trial to the cumulative data of this outcome. TSA suggested 133 meta-analyzed intervention effects to be true beneficial (dark green area) and 133 as potentially false beneficial (light green area). TSA suggested 214 meta-analyzed intervention effects as true neutral (dark yellow area) and 541 as potentially false neutral (light yellow area). TSA suggested 42 meta-analyzed intervention effects as true harmful (dark red area) and 17 as potentially false harmful (light red area). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 1. Proportion of intervention effects on an outcome according to TSA boundaries in primary and sensitivity analyses

	Primary analysis	Sensitivity analyses			Robust conclusions	
Intervention effects according to TSA	**RRR est; D^2 est**	**RRR est; D^2 25%**	**RRR 25%; D^2 est**	**RRR 25%; D^2 25%**	**Meta-analyses crossing the TSA boundary at least in one of TSA analyses**	**Meta-analyses crossing TSA boundaries in all TSA analyses**
True positive	133 (12.3%)	138 (12.8%)	112 (10.4%)	116 (10.7%)	216 (20.0%)	43 (4.0%)
False positive	133 (12.3%)	132 (12.2%)	160 (14.8%)	157 (14.5%)		
True neutral	214 (19.8%)	206 (19.1%)	243 (22.5%)	256 (23.7%)	216 (20.0%)	150 (13.9%)
Potentially false neutral	541 (50.1%)	548 (50.7%)	503 (46.6%)	490 (45.4%)		
True negative	17 (1.6%)	21 (1.9%)	16 (1.5%)	15 (1.4%)	43 (4.0%)	1 (0.01%)
Potentially false negative	42 (3.9%)	35 (3.2%)	46 (4.3%)	46 (4.3%)		
Total	1,080	1,080	1,080	1,080		

Abbreviations: D^2 est, diversity as estimated in the meta-analysis; RRR est, RRR as estimated in the meta-analysis; TSA, Trial Sequential Analysis.
Percentages refer to the total of 1,080 meta-analyzed outcomes.
The range of robust conclusion of meta-analyzed outcomes (independent of potential risks of systematic errors) ranges for true positive findings from 4.0% to 20%, for true negative findings from 0.01% to 4.0% and for true neutral findings from 14% to 20%.

reported nonrandomized trials. Forty articles were not assessed for risk of bias due to insufficient capacity for translation (Chinese $n = 36$; Japanese $n = 2$; Hebrew $n = 1$; and Italian $n = 1$). We assessed that 25 randomized trials (6.7% of 371) of the meta-analyses with a beneficial intervention effect and two randomized trials (11% of 19) of the meta-analyses with a harmful intervention effect were at overall low risk of bias (Supplemental data, Appendix D).

In 95% of all 390 RCTs, the assessors agreed on the rating of the overall risk of bias, whereas in 5.5%, agreement was reached after discussions.

Of the 44 meta-analyses with true beneficial or harmful intervention effects suggested by all four TSAs, there were 10 meta-analyses, which had been evaluated by two or more randomized trials at overall low risk of bias.

One meta-analysis showed a true beneficial intervention effect and one other meta-analysis suggested a true harmful intervention effect according to TSA (Fig. 4).

4. Discussion

We identified 464 reviews of interventions in critical care with 1,080 (co-)primary meta-analyzed outcomes. Assessment of the risks of random errors of these meta-analyses revealed 886 (82%) potentially inconclusive

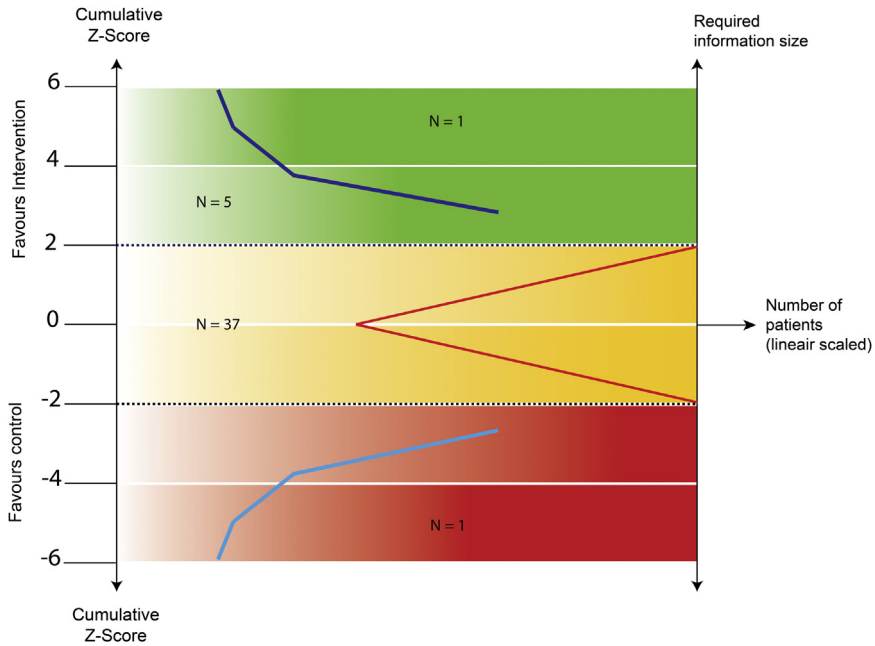

Fig. 4. Results of the analysis including only the overall low risk of bias trials of the meta-analyses, which suggested in the primary analysis either a true beneficial ($N = 43$) or a true harmful ($N = 1$) effect. Only one true beneficial effect remained true beneficial (dark green area) and also one harmful effect remained true harmful (dark red area). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

intervention effect estimates, for example, 281 of 325 conventionally claimed significant effects, and 605 of 755 conventionally claimed neutral effects. Of 266 meta-analyses with a beneficial intervention effect in conventional meta-analysis, 43 (16%) showed an overall ''true'' positive effect in all four TSAs. Of the 59 meta-analyses with harmful intervention effect in conventional meta-analyses, only one (1.7%) actually showed a ''true'' harmful effect in all four TSAs. After assessment of risks of bias of the RCTs included in these true effect meta-analyses and conducting TSA with only the RCTs at low overall risks of bias, one meta-analysis (0.01%) showed a true beneficial effect, and one meta-analysis (0.01%) showed a true harmful effect.

After adjusting for random error using TSA, 87% of seemingly conclusive significant effect estimates of meta-analyses on interventions in critical care appeared in fact inconclusive. When additionally risks of bias of the RCTs were taken into consideration, 323 of 325 (99%) meta-analyses, which suggest a significant effect by conventional meta-analysis, could be considered inconclusive. In addition, 50% of the neutral meta-analyses are potentially false neutral and may eventually appear beneficial (or harmful) once further evidence emerges.

TSA may identify inconclusive results based on realistic assumptions [27,29]. We required the intervention effects to draw the same conclusion in all the primary and sensitivity analyses before considering the intervention effect as ''true'' (4%), which expresses the uncertainties considering the estimates for a realistic RRR and D^2. If we apply less strict criteria, that is, accepting only one of the four TSAs for arriving at a robust conclusion, then 20% (compared with 4.0%) of all meta-analyses crossed the boundary for benefit in at least one of the four TSAs. Still, 56% of all 1,080 meta-analyses do not cross any of the boundaries prohibiting reliable conclusions, so even when using less strict criteria, the majority of meta-analyses have indefinite conclusions. On the other hand, our criteria to consider an intervention as truly harmful might be too rigid. According to the European Medicine Agency's guidelines on multiplicity issues in RCTs, harmful intervention effects should be carefully assessed depending on seriousness, severity, or outcome, irrespective of the *P* value observed [30]. One could argue that when an intervention effect after proper investigation suggests harm irrespective of a conclusive TSA or not, the intervention should not be used.

In systematic reviews, the GRADE approach is often used to assess the quality and hence the certainty of evidence. One of the aspects of this rating is imprecision. GRADE recommends ''in general, results are imprecise when studies include relatively few patients and few events and thus have a wide CI around the estimate of effect.'' Both GRADE and TSA emphasize the importance of the information size required to make judgments on intervention effects presented in systematic reviews. TSA estimates the ''required'' or ''optimal'' information size, whereas GRADE emphasizes its importance without giving specific guidance on judgments.

The choice of a 25% RRR in the TSA analyses is debatable as it may well be an unrealistic high a priori intervention effect, yet it is also suggested by GRADE, although, the GRADE concept of optimal information size does not consider adjusting its size for statistical heterogeneity, making it suboptimal for grading precision in random-effects meta-analyses [31].

According to GRADE, each outcome in systematic reviews is considered separately. However, one could speculate that the type I error risk for each coprimary outcome ought to be reduced for the family-wise error rate (FWER) to remain $<5\%$. This means for this study that reviews with three included (co-)primary outcomes should each use a maximal type I error risk of 0.025 to limit the FWER to 0.05 [18,32]. This would have led to even higher numbers of inconclusive results and larger proportions of potentially false positive (or neutral or negative) conclusions.

The 95% confidence limit is commonly used as the cutoff point where certainty begins. It is worth remembering that the 95% cutoff is itself arbitrary and alternative type I error risks of, for example, 1% have been suggested recently [33,34]. So, the TSA can be regarded as an adjustment of the confidence limits, which follows sound rationale to preserve an overall 5% type I error considering sparse data and repetitive testing in cumulative meta-analyses.

There are several limitations to this meta-epidemiologic study. First, preferably, we should have evaluated the risks of bias of all RCTs included in all meta-analyses and not only in those that showed firm evidence for a beneficial or harmful effect. We may have missed an intervention with a true beneficial effect in which the high-risk studies underestimated the effect estimates. However, this is unlikely as trials at overall high risk of bias are on average associated with overestimation of beneficial effects and underestimation of harmful effects so that one would not expect a given nonbeneficial intervention effect to become beneficial after correction for bias risk assessment [35]. We abstained from bias evaluation of all 8,195 RCTs assessing the 1,080 meta-analyses as this would have required a huge effort with minimal chances of finding a beneficial effect, especially because the power to detect an effect would decrease substantially when including only the trials with overall low risk of bias. Second, we included meta-analyses from both systematic and nonsystematic reviews. It is likely that meta-analyses from systematic reviews contain less risks of random error compared with meta-analyses from nonsystematic reviews. However, the distinction between a systematic and nonsystematic review is not always clear. Authors of reviews who claim to have conducted a systematic review frequently appear to have major flaws in design and/or conduct, violating the criteria associated with the highest level of evidence. In fact, the distinction between

systematic and nonsystematic is not dichotomous but rather a continuum (or more or less a categorical variable) with increasing levels of alignment fulfilling the highest level of evidence.

Third, we intended to include all studies without any language restrictions. Unfortunately, we had insufficient capacity for all translations so that we were forced to exclude 25 meta-analyses from our analyses. In addition, we also assumed 40 RCTs included in meta-analyses with a true beneficial intervention effect to be at high risk of bias due to insufficient capacity for translations.

Fourth, we have used the DerSimonian-Laird random effects estimator, although newer methods have been found to have improved performance [36]. The DerSimonian-Laird method is the most widely used method and is also programmed in TSA software, which is why we used this method.

5. Conclusions

There are large risks of random errors associated with meta-analyses in critical care: at least 50% of all meta-analyses that claim a significant intervention effect based on conventional statistical methods are potentially false positive. Also, 80% of all meta-analyses on interventions which claim a nonsignificant intervention effect based on conventional statistical methods may be potentially false neutral. Thus, some interventions in critical care may be falsely accepted, whereas many others may be falsely rejected in daily practice. Future studies, in critical care, should aim for establishing an effect of interventions accounting for risks of bias and risks of random errors.

6. Deviations from protocol

In our protocol, we stated that some interventions cannot be blinded for either participants or caregivers. For the meta-analyses of these interventions, we would classify trials at low risk of bias trials if all domains except blinding of participant and personnel were at low risk of bias. Eventually, we decided not to do this since judgments on whether interventions can or cannot be blinded are frequently disputable. Also, irrespective whether interventions can or cannot be blinded, there will always be the risk of bias when not using blinding an intervention.

Also, in our protocol, we stated that in the case of zero events, we would make an empirical adjustment of 0.001 to the number of events in the control and intervention groups. However, we used 0.01 because the TSA program used adjustment of 0.01 for zero events.

Last, we did not asses the risk of bias of the domains of vested interest bias, and any other bias risk as suggested by the Cochrane risk of bias tool. We did not assess these risks of bias because of its complexity and need for thorough background check. Accordingly, the risks of bias may be worse than assessed by us.

CRediT authorship contribution statement

Thijs M. Koster: Writing - original draft, Writing - review & editing. **Jørn Wetterslev:** Writing - original draft, Writing - review & editing. **Christian Gluud:** Writing - original draft, Writing - review & editing. **Janus C. Jakobsen:** Writing - review & editing. **Thomas Kaufmann:** Writing - review & editing. **Ruben J. Eck:** Writing - review & editing. **Geert Koster:** Writing - review & editing. **Bart Hiemstra:** Writing - review & editing. **Iwan C.C. van der Horst:** Writing - review & editing. **Eric Keus:** Writing - original draft, Writing - review & editing.

Acknowledgments

The authors wish to thank Sarah Klingenberg (Copenhagen Trial Unit) for her help with the search strategy.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.05.011.

References

[1] Ioannidis JP. Meta-analyses can be credible and useful: a new standard. JAMA Psychiatry 2017;74:311−2.
[2] Mulrow CD. Rationale for systematic reviews. BMJ 1994;309:597−9.
[3] Pogue J, Yusuf S. Overcoming the limitations of current meta-analysis of randomised controlled trials. Lancet 1998;351:47−52.
[4] Ospina-Tascón GA, Büchele GL, Vincent J. Multicenter, randomized, controlled trials evaluating mortality in intensive care: doomed to fail? Crit Care Med 2008;36:1311−22.
[5] Vincent J. We should abandon randomized controlled trials in the intensive care unit. Crit Care Med 2010;38:S534−8.
[6] Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions Version 5.1.0, The Cochrane Collaboration. 2011.
[7] Koster T, Wetterslev J, Gluud C, Keus F, van der Horst I. Systematic overview and critical appraisal of meta-analyses of interventions in intensive care medicine. Acta Anaesthesiol Scand 2018;62:1041−9.
[8] Humaidan P, Polyzos NP. (Meta)analyze this: systematic reviews might lose credibility. Nat Med 2012;18:1321.
[9] Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. J Clin Epidemiol 2008;61:64−75.
[10] Wetterslev J, Jakobsen JC, Gluud C. Trial sequential analysis in systematic reviews with meta-analysis. BMC Med Res Methodol 2017;17:39.
[11] Gordon Lan KK, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika 1983;70:659−63.
[12] Thorlund K, Devereaux P, Wetterslev J, Guyatt G, Ioannidis JP, Thabane L, et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? Int J Epidemiol 2008;38:276−86.
[13] Devereaux PJ, Beattie WS, Choi PT, Badner NH, Guyatt GH, Villar JC, et al. How strong is the evidence for the use of perioperative beta blockers in non-cardiac surgery? Systematic review and meta-analysis of randomised controlled trials. BMJ 2005;331:313−21.
[14] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009;151:264−9.

[15] Whiting P, Savović J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol 2016;69:225−34.
[16] Phillips B, Ball C, Badenoch D, Straus S, Haynes B, Dawes M. Oxford centre for evidence-based medicine levels of evidence (May 2001). BJU Int 2011;107:870.
[17] Pussegoda K, Turner L, Garritty C, Mayhew A, Skidmore B, Stevens A, et al. Systematic review adherence to methodological or reporting quality. Syst Rev 2017;6:131.
[18] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924−6.
[19] Turner RM, Bird SM, Higgins JP. The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. PLoS One 2013;8:e59202.
[20] 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.
[21] Kim K, DeMets DL. Confidence intervals following group sequential tests in clinical trials. Biometrics 1987;43:857−64.
[22] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177−88.
[23] O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35:549−56.
[24] Jennison C, Turnbull BW. Group sequential methods with applications to clinical trials. Boca Raton: Chapman and Hall/CRC; 1999:416.
[25] Sweeting M J, Sutton A J, Lambert P C. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat Med 2004;23:1351−75.
[26] Higgins JPT, Thompson S. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21:1539−58.
[27] 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.
[28] Thorlund K, Imberger G, Johnston BC, Walsh M, Awad T, Thabane L, et al. Evolution of heterogeneity (I2) estimates and their 95% confidence intervals in large meta-analyses. PLoS One 2012;7:e39471.
[29] 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 2008;38:287−98.
[30] European Medicines Agency. Multiplicity issues in clinical trials, European Medicines Agency 2017: https://www.ema.europa.eu/en/multiplicity-issues-clinical-trials.
[31] Gartlehner G, Nussbaumer-Streit B, Wagner G, Patel S, Swinson-Evans T, Dobrescu A, et al. Increased risks for random errors are common in outcomes graded as high certainty of evidence. J Clin Epidemiol 2019;106:50−9.
[32] Jakobsen JC, Wetterslev J, Lange T, Gluud C. Taking into account risks of random errors when analysing multiple outcomes in systematic reviews. Cochrane Database Syst Rev 2016ED000111.
[33] Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nat Hum Behav 2018;2(1):6−10.
[34] Ioannidis JPA. The importance of predefined rules and prespecified statistical analyses: do not abandon significance. JAMA 2019;321:2067−8.
[35] Savović J, Jones HE, Altman DG, Harris RJ, Jüni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med 2012;157:429−38.
[36] 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.