Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2009;**38**:287–298
doi:10.1093/ije/dyn188

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

Jesper Brok,* Kristian Thorlund, Jørn Wetterslev and Christian Gluud

Accepted 13 August 2008

Background Random error may cause misleading evidence in meta-analyses. The required number of participants in a meta-analysis (i.e. information size) should be at least as large as an adequately powered single trial. Trial sequential analysis (TSA) may reduce risk of random errors due to repetitive testing of accumulating data by evaluating meta-analyses not reaching the information size with monitoring boundaries. This is analogous to sequential monitoring boundaries in a single trial.

Methods We selected apparently conclusive ($P \leqslant 0.05$) Cochrane neonatal meta-analyses. We applied heterogeneity-adjusted and unadjusted TSA on these meta-analyses by calculating the information size, the monitoring boundaries, and the cumulative *Z*-statistic after each trial. We identified the proportion of meta-analyses that did not reach the required information size and the proportion of these meta-analyses in which the *Z*-curve did not cross the monitoring boundaries.

Results Of 54 apparently conclusive meta-analyses, 39 (72%) did not reach the heterogeneity-adjusted information size required to accept or reject an intervention effect of 25% relative risk reduction. Of these 39, 19 meta-analyses (49%) were considered inconclusive, because the cumulative *Z*-curve did not cross the monitoring boundaries. The median number of participants required to reach the required information size was 1591 (range, 339–6149). TSA without heterogeneity adjustment largely confirmed these results.

Conclusions Many apparently conclusive Cochrane neonatal meta-analyses may become inconclusive when the statistical analyses take into account the risk of random error due to repetitive testing.

The Copenhagen Trial Unit, Center for Clinical Intervention Research, Department 3344, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark
* Corresponding author: The Copenhagen Trial Unit, Center for Clinical Intervention Research, Department 3344, Rigshospitalet, Copenhagen University Hospital, Blegdamsvej 9, DK-2100 Copenhagen, Denmark. E-mail: jbrok@ctu.rh.dk

Keywords Meta-analysis, trial sequential analysis, random error, information size, sample size, heterogeneity

Introduction

Meta-analyses of randomized trials are considered the gold standard for intervention comparisons.[1–3] However, meta-analyses are not errorless. Random error ('play of chance') is one reason for misleading results in meta-analyses.[4,5] The risk of random error may increase considerably due to multiple looks on accumulating evidence when new trials emerge.[4,5]

The standards for testing statistical significance in meta-analyses should be, at least, equal to those of a randomized trial.[6–9] In a randomized trial it is essential to perform an a priori sample size estimation. By the same token, a meta-analysis should include an information size at least as large as the sample size of an adequately powered single trial to reduce the risk of random error.[6–9] Given the greater risk for additional biases and heterogeneity among trial designs in a meta-analysis compared with a single randomized trial, even more information is likely to be required in a meta-analysis.[8,9] Despite these simple prerequisites the medical communities have largely ignored the issues of information size and risks of random errors in meta-analyses.

The aim of a meta-analysis is to identify the benefit or harm of an intervention as early as possible. Thus, meta-analyses are commonly updated when new trials are published, i.e. repeated analyses are performed on accumulating data.[1,2] Such multiple looks induce repeated significance testing, which, if performed with the conventional *P*-value criterion (typically two-sided $\alpha = 5\%$), is prone to exacerbate the risk of random error.[4,5] The situation is comparable to interim analyses of a single randomized clinical trial. In clinical trials conservative adjustments are commonly made to the thresholds for declaring one intervention significantly better than another.[10] Such adjustments may be performed through the use of formal sequential monitoring boundaries that function as a threshold for the employed test statistic. Typically, sequential monitoring boundaries in a single clinical trial demand a conservative interpretation when data are sparse, but become increasingly lenient as more data accumulate. Similar utilization of formal boundaries as guides for cumulative meta-analyses is desirable to distinguish real effects from random errors.[9]

Trial sequential analysis (TSA) is a methodology that combines an a priori information size calculation for a meta-analysis with the adaptation of monitoring boundaries to evaluate the accumulating data (i.e. meta-analytic updates).[9,11] The information size calculation is similar to the sample size calculation in a single trial, which (for binomial data) requires an a priori realistic event proportion in the control group, a minimal intervention effect size that is relevant or judged clinically worthwhile and biologically plausible, and a desired maximum risk of statistical errors (usually with $\alpha = 0.05$ and $\beta = 0.2$).[9,11] Once the information size is calculated, trial sequential monitoring boundaries can be adapted as new trials are published and meta-analyses are updated over time. In this context, TSA may serve as a tool for quantifying the reliability of cumulative data in meta-analyses.[9,11]

In this study we identified all apparently conclusive Cochrane Neonatal Group systematic reviews that recommend an intervention based on at least one meta-analysis with a *P*-value $\leqslant 0.05$.[12,13] We applied TSA on these meta-analyses, i.e. we calculated the required information size and constructed the trial sequential monitoring boundaries. We explored the extent to which apparently conclusive meta-analyses remained conclusive when accounting for potentially exacerbated risk of random error due to repetitive testing. We utilized TSA to evaluate the risk of random error, and calculated the additional required information to ascertain conclusiveness. Further, we discuss the relative merits and pitfalls of the application of TSA to meta-analyses, both in a clinical and methodological context.

Methods

Material

We selected apparently conclusive meta-analyses assessing a binary outcome with a *P*-value $\leqslant 0.05$ from the Cochrane Neonatal Group reviews recommending an intervention (The Cochrane Library, Issue 4, 2004).[12,13]

Trial sequential analyses

TSA necessitates pre-specification of a relevant (worthwhile) intervention effect (μ) and risk of type 1 (α) and type 2 (β) errors.[9,11] We set two-sided $\alpha = 5\%$ and $\beta = 20\%$ ($1 - \beta = 80\%$ power). The required information size was calculated using the formula

$$2 \cdot (Z_{\alpha/2} + Z_{\beta})^2 \cdot 2 \cdot v/\mu^2.$$

Here $\mu = P_C - P_E$ denotes the intervention effect (P_C and P_E being the proportion in the control group and in the intervention group with the outcome) and $v = P^* \cdot (1 - P^*)$ its variance assuming $P^* = (P_C + P_E)/2$, i.e. equal size of the intervention and the control group. We estimated P_C by meta-analyzing the control group event proportions of all included trials. Using a 25% relative risk reduction and the estimated control group event proportion we obtained

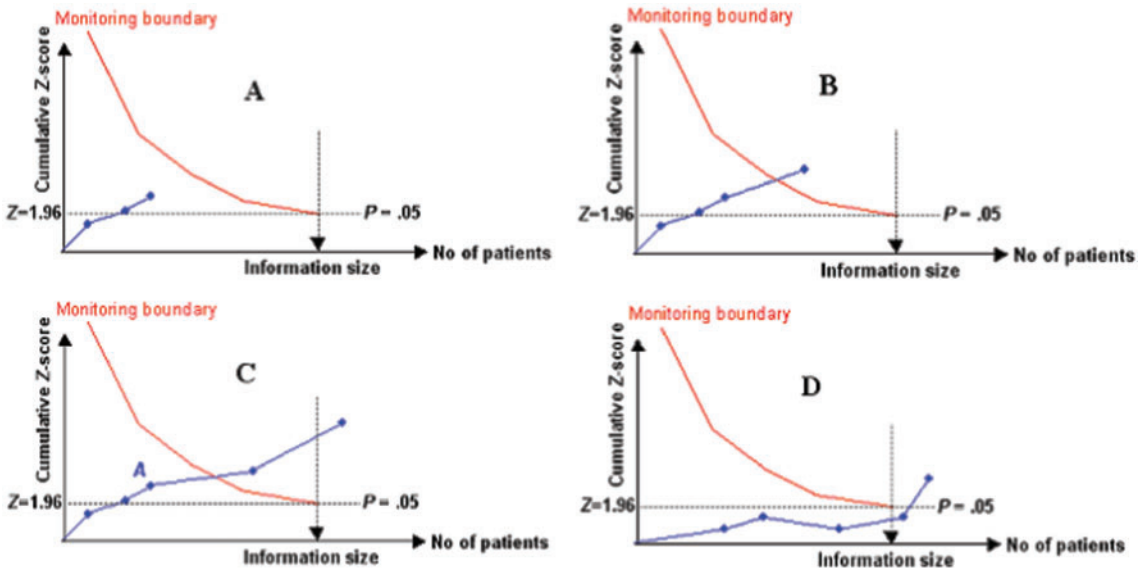


Figure 1 Four examples of TSA. The cumulative *Z*-curves (blue) were constructed with each cumulative *Z*-value calculated after including a new trial according to publication date. Crossing of the two-sided $Z = 1.96$ provides a traditionally significant result. Crossing of the trial sequential monitoring boundaries (red) is needed to obtain reliable evidence. (A) Inconclusive evidence: Number of participants does not reach the information size and the cumulative *Z*-curve does not cross the monitoring boundary. (B) Evidence for at least 25% relative risk reduction: Number of participants does not reach the information size, but the cumulative *Z*-curve does cross the monitoring boundary. (C) Evidence for at least 25% relative risk reduction: Number of participants does reach the information size and the cumulative *Z*-curve does cross the monitoring boundary. (D) Evidence of less than 25% relative risk reduction: The cumulative *Z*-curve does not cross the monitoring boundary before reaching the information size

the experimental intervention group event proportion, P_E. We also conducted sensitivity analyses assuming a relative risk reduction of 15%.

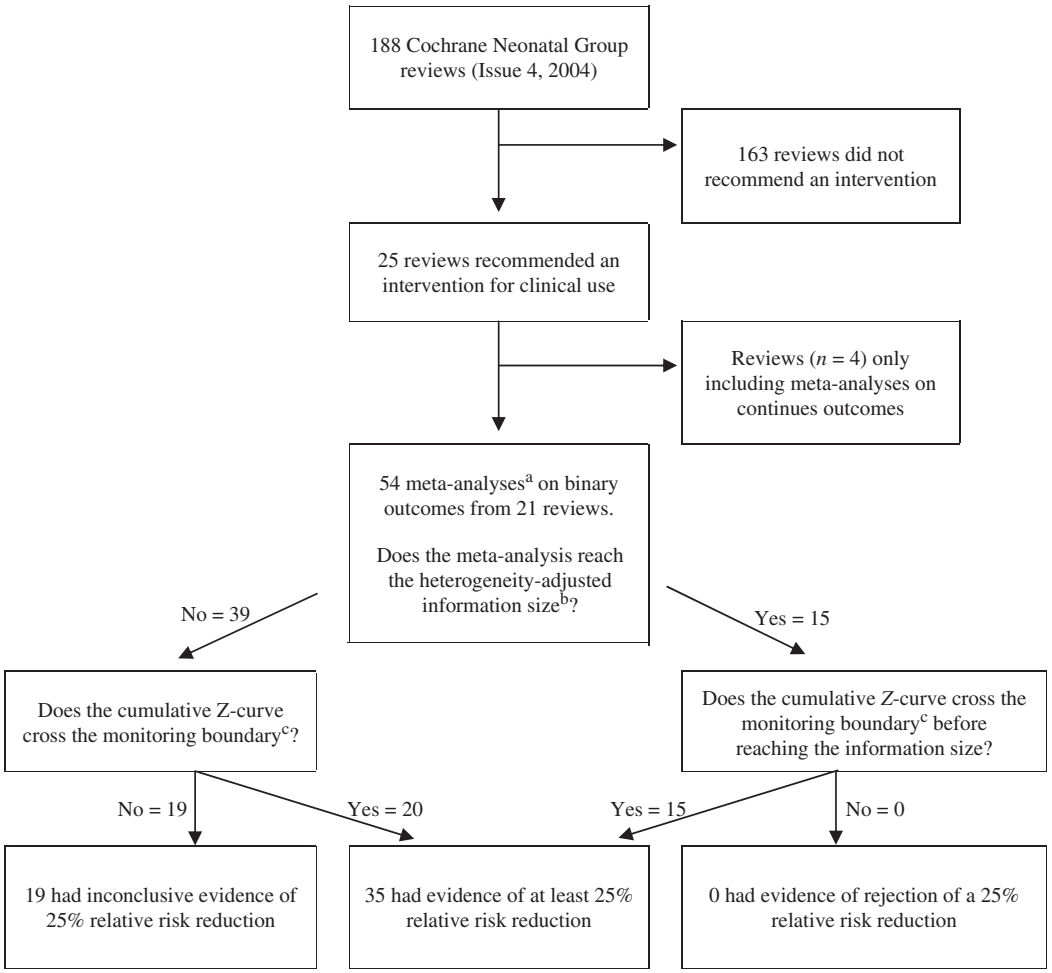
Heterogeneity increases the uncertainty in a meta-analyses.[14] Heterogeneity may be measured by I^2.[14] We adjusted the required information size according to the degree of heterogeneity expressed by I^2 by multiplying the required information size (see above) by $1/(1-I^2)$.[9] This may correspond to the heterogeneity adjustment in a multi-centre trial.[15] As a sensitivity analysis, we conducted TSA without the heterogeneity-adjustment as all analyses were originally conducted as fixed-effect model analyses (see below).

For each meta-analysis we calculated the heterogeneity-adjusted and unadjusted information size as described and applied the trial sequential monitoring boundaries.[9] The monitoring boundaries were based on the Lan–DeMets α-spending function that controls the overall type I error by spending it in an appropriate manner, as statistical tests are employed throughout the accumulation of trials.[10,16] We choose the α-spending function that results in the well-known O'Brien–Fleming monitoring boundaries.[17] We calculated the cumulative *Z*-curve of each cumulative meta-analysis (i.e. the series of *Z*-statistics after each consecutive trial) and assessed its crossing of monitoring boundaries with the fixed-effect model[18] or random-effects model[19] as used in the Cochrane review. The monitoring boundaries should be crossed by the cumulative *Z*-curve to obtain firm evidence for an intervention effect (Figure 1). *Z*-values of ± 1.96 correspond to the conventional $P = 0.05$ in a two-sided hypothesis test.

Data (title, author, publication year, intervention(s), outcome, number with the outcome in question in the intervention and control group, number of participants in the intervention and control group, and number of trials) from each meta-analysis was extracted by one author (JB). Data was analysed with our Copenhagen Trial Unit computer program, TSA v0.8. Correct data extraction and entry were verified by comparing the final *Z*-score for each meta-analysis in TSA v0.8 with *Z*-score obtained in Review Manager 4.10, which is the standard program used by Cochrane review authors.[20] If identical, correct data extraction and entry were assumed. The TSA v0.8 displayed the relationship between the cumulative *Z*-score, the information size, and the two-sided monitoring boundaries on a graph. For simplicity we only show the monitoring boundary regarding benefit (or harm) of the experimental intervention.

Outcomes

The proportion of meta-analyses that did not reach the information size, and the proportion of such meta-analyses in which the cumulative *Z*-curve did not cross the monitoring boundary were calculated (Figure 1A). For such meta-analyses we calculated the additional number of participants required to reach the required information size. Meta-analyses



[a] $P \leq 0.05$.
[b] Participants required to accept (or reject) a predefined intervention effect of 25% adjusted for heterogeneity.
[c] Constructed according to an a priori intervention effect of 25% relative risk reduction.

Figure 2 Flowchart for selection of meta-analyses and subsequent trial sequential analyses of 54 apparently conclusive ($P \leqslant 0.05$) Cochrane neonatal meta-analyses

that crossed the monitoring boundary show evidence of an intervention effect of at least 25% relative risk reduction (Figure 1B).

For meta-analyses that reached the required information size we identified the proportion that crossed the monitoring boundary before reaching the information size, which show evidence for an intervention effect of at least 25% relative risk reduction (Figure 1C). The remaining meta-analyses rejected an intervention effect of 25% relative risk reduction (Figure 1D).

Results

We identified 25 out of the 188 Cochrane Neonatal Group reviews in The Cochrane Library, Issue 4, 2004 that recommended an intervention for clinical use (Figure 2).[12,13] Of these, we excluded four (16%) because they did not report a meta-analyses on a binary outcome measure with *P*-values $\leqslant 0.05$. From the remaining 21 reviews we included 54 meta-analyses demonstrating a beneficial intervention effect (i.e. *P*-value $\leqslant 0.05$).[21–41]

These meta-analyses included a median of five randomized trials (range 1–14) and a median of 932 participants (range 32–4588). The Cochrane authors analyzed all meta-analyses with relative risks according to the fixed-effect model and compared the experimental intervention vs placebo or no intervention ($n = 30$) or vs another intervention ($n = 24$). The final *Z*-values ranged from -6.22 to -1.99 (*P*-values <0.00001 to 0.05).

Heterogeneity-adjusted information size in the 54 meta-analyses

The accrued number of participants did not reach the required information size in 39 meta-analyses (72%) required to accept or reject an intervention effect of 25% relative risk reduction (Figure 2).

Evidence in 39 meta-analyses, which did not reach the heterogeneity-adjusted information size

Figure 2 and Table 1 summarize the overall findings. The cumulative *Z*-curve did not cross the monitoring boundary in 19 meta-analyses (49%), which did not reach the heterogeneity-adjusted information size (Figure 1A). In these meta-analyses, the median additional information size required to obtain evidence for or against 25% relative risk reduction was 1591 participants (range 339–6149) (Table 1 and Figure 3). The cumulative *Z*-curve crossed the monitoring boundary in the remaining 20 meta-analyses (51%), showing evidence for an intervention effect of at least 25% relative risk reduction (Figure 1B).

Evidence in 15 meta-analyses, which reached the heterogeneity-adjusted information size

The cumulative *Z*-curve crossed the monitoring boundary showing evidence for an intervention effect of at least 25% relative risk reduction in all 15 meta-analyses (100%), which reached the heterogeneity-adjusted information size (Figure 1C).

Sensitivity analyses calculating information size without heterogeneity adjustment

The accrued number of participants did not reach the required information size in 36 meta-analyses (67%) (Figure 2). Of these, the cumulative *Z*-curve did not cross the monitoring boundary in 13 meta-analyses (36%) (Figure 1A).

The cumulative *Z*-curve crossed the monitoring boundary in all 18 meta-analyses (100%) that did reach the required information size showing evidence for an effect of at least 25% relative risk reduction (Table 1; Figure 1C).

Sensitivity analyses based on 15% relative risk reduction with heterogeneity adjustment

The accrued number of participants did not reach the required information size in 50 meta-analyses (93%). Of these, the cumulative *Z*-curve did not cross the monitoring boundary in 36 meta-analyses (72%) (Figure 1A).

The cumulative *Z*-curve crossed the monitoring boundary showing evidence for an effect of at least 15% relative risk reduction in all four meta-analyses (100%) that reached the required information size (Figure 1C).

Discussion

Our assessments showed that three out of four neonatal meta-analyses considered conclusive have insufficient heterogeneity-adjusted information size to accept (or reject) a 25% relative risk reduction. Reanalyzing these meta-analyses with trial sequential monitoring boundaries revealed that almost half of the meta-analyses had inconclusive evidence if adjusted for the risk of random error (Table 1). On average, it may be necessary to acquire additional 1600 participants in order to obtain sufficient evidence for or against a 25% relative risk reduction.

Why use trial sequential analysis?

The prevalence of meta-analyses at risk of random error due to repetitive testing seems too high to be ignored.[4–9,11,42] TSA retains the desired risk of random error when repeated conventional significance testing on accumulating data in cumulative meta-analysis are performed.[9,11] While it may be argued that no adjustment is needed if a meta-analysis is only carried out once,[4] 18% of meta-analyses (systematic reviews) are reported as being updates[43] and meta-analyses in Cochrane reviews should be updated when valid new evidence emerges or at least every second year.[2] The meta-analysis paradigm is still in its infancy and the yearly number is still increasing. Furthermore, trial authors are also encouraged to carry out meta-analyses before and after the conduct of a new trial.[1,44] Moreover, we do not know the frequency with which health-care workers are doing meta-analyses without publishing them because they may be waiting for an 'interesting' result. This represents a weakness of meta-analyses, i.e. the retrospective nature of the research process.[2] Hence, repetitive testing probably exists in most meta-analyses.

Theoretical considerations suggest that *cumulative* meta-analyses of sparse data have a substantial probability of overestimating effects due to random variation.[4,5] Such considerations are also supported by empirical evidence.[6–9,11] In the context of repetitive testing, a meta-analysis of sparse evidence will typically be among the first in a series of meta-analytic updates. Thus, enforcing some degree of conservatism at this stage seems appropriate. We have achieved this in our study by employing the Lan–DeMets α-spending function using the O'Brien–Fleming monitoring boundaries.[10,16,17] As authors of meta-analyses focus primarily on the point estimates and confidence intervals such conservatism should also be employed to control the desired coverage level of the confidence intervals. Methods for controlling coverage levels in the setting of repetitive testing have been developed.[11]

Obviously, TSA provides a more conservative conclusion, which may delay clinicians' use of an intervention.[45] This conflicts with the current 'societal opinion' that clinicians find it difficult to do nothing and that errors of omission are considered more reprehensible than errors of commission.[46] Such delay should be weighted against the risk of introducing interventions in which the benefit-risk ratio is based on inconclusive evidence.[47–49] In this vein, adjusted

Table 1 Apparently conclusive[a] Cochrane neonatal meta-analysis considered to be inconclusive because the required information size[b] was not reached and the cumulative *Z*-curve did not cross the monitoring boundary

Review	Meta-analysed outcome	Intervention(s)	Number. of patients (Trials)	Relative risk reduction (95% CI)	I^2	*Z*-score (*P*-value)	Heterogeneity-adjusted information size[b]	Unadjusted information size[b]	Maximum additional patients required
(A) Puckett *et al.*[21]	Bleeding	Prophylactic vitamin K	3338 (1)	0.27 (0.04–0.44)	0%	−2.26 (0.02)	5445	5445	2107
(B) Soll *et al.*[22]	Pneumothorax	Prophylactic synthetic surfactant	1252 (6)	0.33 (0.10–0.50)	43%	−2.68 (0.007)	3273	1840	2021
(C) Soll *et al.*[22]	Mortality	Prophylactic synthetic surfactant	1046 (3)	0.17 (0.02–0.30)	56%	−2.14 (0.03)	1550	682	504
(D) Osborn *et al.*[23]	Asthma	Formulas with hydrolysed protein	945 (6)	0.41 (0.14–0.60)	0%	−2.72 (0.006)	2016	2016	1071
(E) Fowlie *et al.*[24]	Periventricular leukomalacia	Prophylactic indomethacin	811 (5)	0.56 (0.19–0.86)	0%	−2.64 (0.008)	3998	3998	3187
(F) Yost *et al.*[25]	Pulmonary emphysema	Early vs delayed surfactant	737 (2)	0.37 (0.07–0.57)	0%	−2.30 (0.02)	1873	1873	1136
(G) Bell *et al.*[26]	Mortality	Restricted vs liberal water intake	414 (4)	0.48 (0.04–0.72)	41%	−2.09 (0.04)	3964	2345	3550
(H) Soll *et al.*[27]	Pneumothorax	Multiple vs single dose surfactant	394 (2)	0.49 (0.12–0.70)	0%	−2.45 (0.01)	2058	2058	1664
(I) Henderson-Smart *et al.*[28]	Mortality	Mechanical ventilation	359 (5)	0.14 (0.01–0.26)	54%	−1.99 (0.05)	394	182[d]	45
(J) Bell *et al.*[9]	Patent ductus arteriosus	Restricted vs liberal water intake	358 (3)	0.60 (0.37–0.74)	51%	−4.06 (<0.0001)	1799	876[d]	1441
(K) Bell *et al.*[29]	Necrotizing enterocolitis	Restricted vs liberal water intake	358 (3)	0.70 (0.29–0.87)	61%	−2.74 (0.006)	6149	2383	5791
(L) Ho *et al.*[29]	Mortality	Continuous distending airway pressure	197 (5)	0.48 (0.13–0.68)	0%	−2.51 (0.01)	572	572	375
(M) Puckett *et al.*[21]	Vitamin K deficiency	Prophylactic vitamin K	118 (3)	0.60 (0.34–0.79)	73%	−4.23 (<0.00001)	707	494[d]	589
(N) Steer *et al.*[30]	Adverse events	Caffeine vs theophylline	66 (3)	0.83 (0.28–0.96)	0%	−2.42 (0.02)	863	863	797
(O) Barrington *et al.*[31]	Aortic thrombosis	Low vs high position of umbilical artery catheters	62 (1)	0.69 (0.14–0.89)	0%	−2.24 (0.02)	715	715	653
(P) Barrington *et al* 1999[32]	Aortic thrombosis	End vs side hole of umbilical artery catheters	62 (1)	0.73 (0.33–0.89)	0%	−2.80 (0.005)	535	535	473
(Q) Puckett *et al.*[21]	Vitamin K deficiency	Prophylactic vitamin K	58 (2)	0.57 (0.29–0.74)	63%	−3.31 (0.0009)	339	124[d]	281

(continued)

Table 1 Continued

Review	Meta-analysed outcome	Intervention(s)	Number of patients (Trials)	Relative risk reduction (95% CI)	I^2	Z-score (*P*-value)	Heterogeneity-adjusted information size[b]	Unadjusted information size[b]	Maximum additional patients required
(R) Henderson-Smart *et al.*[33]	Apnea	Continuous positive airway pressure vs theophylline	32 (1)	−1.89[c] (−0.12, −6.47)	0%	2.20 (0.03)	1591	1591	1559
(S) Henderson-Smart *et al.*[33]	Respiratory failure	Continuous positive airway pressure vs theophylline	32 (1)	−2.09[c] (−0.42, −5.70)	0%	2.85 (0.004)	1200	1200	1168

The capitalized alphabet refers to the TSA-panels shown in Figure 3.
[a]$P \leqslant 0.05$.
[b]Participants required to detect (or reject) a predefined intervention effect of 25% with or without adjustment for heterogeneity.
[c]Negative relative risk reduction corresponds to an increased risk.
[d]The cumulative *Z*-curve crossed the monitoring boundary, which provides evidence of at least 25% relative risk reduction.

confidence intervals elucidate the risk of error of commission.

From an investigational perspective TSA provides a quantification and visual overview. In case TSA cannot confirm that conclusive evidence exists, TSA may serve as a valuable tool to estimate what extra efforts that are needed to be able to accept or reject a certain intervention effect.

Strengths and limitations

This study represents the first application of TSA on a large cohort of apparently conclusive meta-analyses. We only evaluated meta-analyses from the Cochrane Neonatal Group. It can be argued that neonatology may show different results compared with other specialties due to the particular ethical considerations within the field of pediatric clinical research and the high child mortality. Thus TSA needs to be applied to meta-analyses within other specialties to confirm if our findings have external validity.

We have only audited meta-analyses of binary outcomes. Meta-analyses on continuous outcomes (e.g. weight and height) are frequently included in neonatal reviews although they often represent weaker surrogate outcomes rather than more patient-important clinical outcomes.[50] As review authors rarely recommend interventions solely on continuous outcomes we excluded only the least important meta-analyses in our study.

Meta-analyses originating from the same review are not independent and are likely to be correlated. This could bias our results. However, using only a single meta-analysis from each Cochrane review, provided the same overall results. We based all TSA on an arbitrary relative risk reduction of 25% and a more conservative one of 15%. Thus, our findings are valid only under the assumption of these effect sizes. Every individual meta-analysis is unique and a smaller or larger pre-specified intervention effect could be more relevant. However, larger effect sizes are rare and a pre-specified assumption of a smaller effect size would make the TSA even more restrictive. Furthermore, use of lower α and β values (e.g. $\alpha = 1\%$ and $\beta = 10\%$) could be considered, but would also induce more conservative TSA.

Our TSA were constructed based on the assumption that a meta-analysis was conducted after each trial. Thus, it accounts for the 'worst case', i.e. a meta-analytic look had been conducted following publication of each trial in the meta-analysis. TSA can also be constructed less conservatively according to the exact number of previous looks (or updates) in a meta-analysis. We used the most restrictive approach, as the number of previous looks in a meta-analysis can be difficult to establish. The difference between these two approaches is, however, negligible. This is due to the mathematical properties of the α-spending function that results in the O'Brien–Fleming monitoring boundaries.[10,16,17] Here, the α-spending occurs

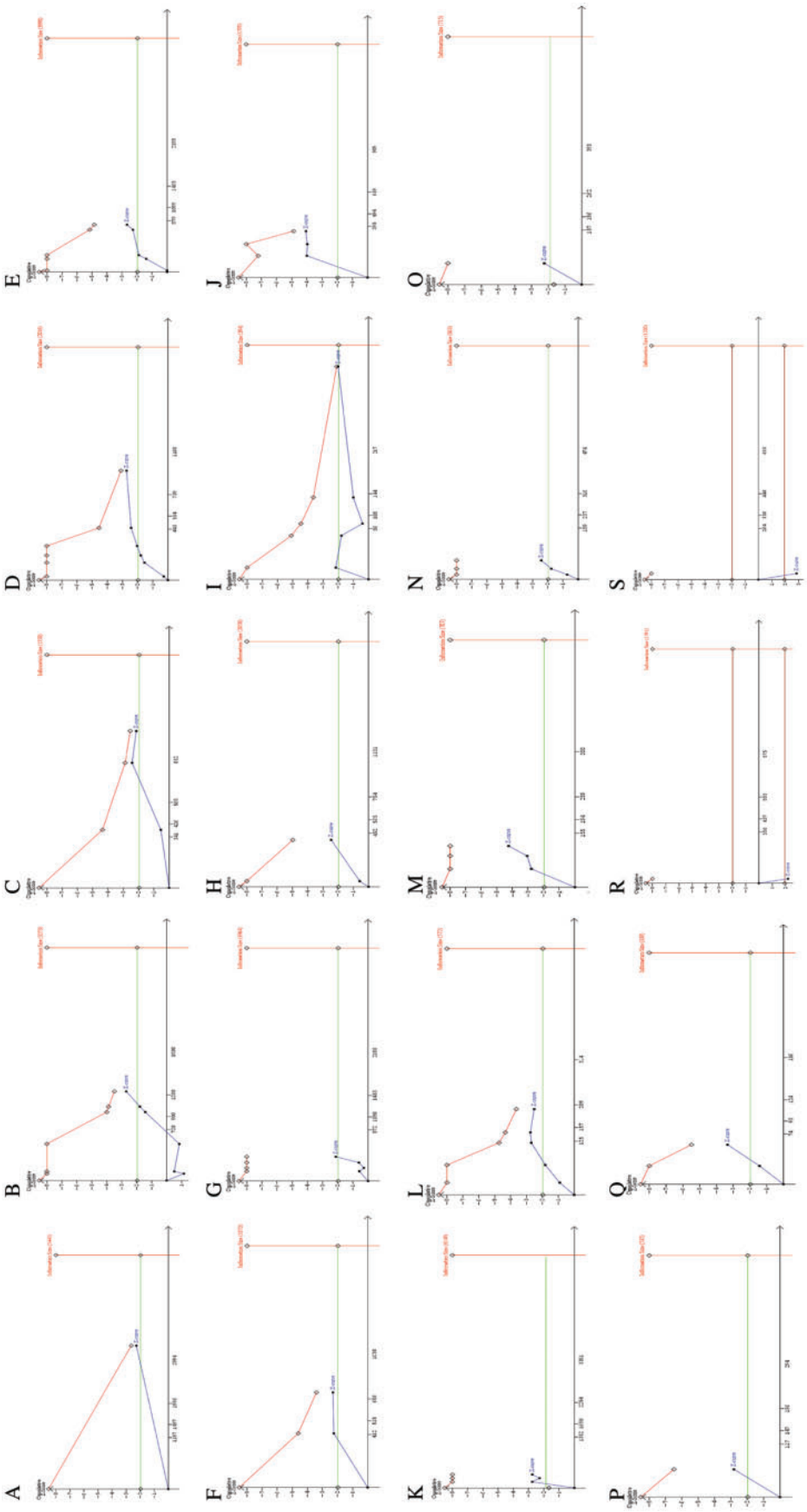


Figure 3 TSA panels of 19 meta-analyses listed in Table 1. The required information size was not reached and the cumulative *Z*-curve did not cross the monitoring boundaries

exponentially to the increment of accumulating patients, and thus, ensures that trial sequential monitoring boundaries are largely insensitive to the number of interim looks conducted before reaching the required information size.

We adjusted TSA for heterogeneity similar to the adjustment for heterogeneity in an individual multicentre trial.[9,11,15] From a meta-analytic perspective it seems important to incorporate the uncertainty related to meta-analyzing trials which may differ regarding populations, interventions, follow-up, etc. In a 'traditional' meta-analysis such heterogeneity may be addressed by using the random-effects model that gives wider confidence intervals when heterogeneity increases.[19] Noteworthy is that all the neonatal reviews used the fixed-effect model meta-analysis irrespective of the heterogeneity level. We are aware that applying heterogeneity-adjusted TSA on meta-analyses analysed with the fixed-effect model may be seen as a violation of the underlying model assumptions, which ignores heterogeneity. Despite not being in concordance with the model chosen, it appears appropriate to incorporate heterogeneity under some form as complete homogeneity across trial populations in a meta-analysis seems unrealistic. Even in meta-analyses without detected heterogeneity, substantial heterogeneity may be present.[51] However, our choice of correcting the information size with the estimated heterogeneity represents only one among many possible adjustments. Our sensitivity analyses without heterogeneity adjustment provided less-conservative results, but still illustrate that a substantial number of meta-analyses may be inconclusive. The difference between TSA with and without heterogeneity adjustment and the difference between fixed- and random-effects models simply highlights the need to carefully consider heterogeneity irrespective of the meta-analytic model.

TSA focuses only on random error in meta-analyses. Systematic error (bias) is also important to assess.[52–57] Several known bias risks in meta-analyses should be considered but studies have indicated that meta-analysts often fail to do so.[58] Furthermore, meta-analyses may be biased due to selective reporting of outcomes.[59] Thus, combining our random-risk correction with appropriate bias-risk adjustments would probably make the number of conclusive meta-analyses supporting an intervention even smaller than observed in our present study.

Relation to similar studies

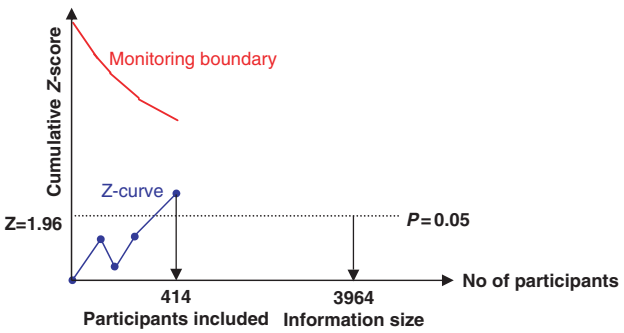
Updated or cumulative meta-analyses are naturally viewed in a Bayesian framework because this revises the information in light of new information.[4,60] From a frequentistic perspective, however, repeated testing increases the risks of overall type I error.[4,6] We took a frequentistic approach that aimed to reduce the type I error. Without appropriate adjustments, even without a genuine treatment effect, adding trials and multiple looks will eventually lead to 'conclusive' evidence.[4,5] Other similar adjustments for multiple looks and heterogeneity in meta-analyses have been suggested based on extensive simulation with use of the 'law of iterated logarithm'.[61,62] The different frequentistic methods should be compared, but the bottom line is that adjustment of meta-analytic evidence seems appropriate. Our present study confirms and extends our previous results after applying TSA on 174 meta-analyses irrespective of the meta-analytic result.[63] In this study we observed that many meta-analyses had insufficient information size and that conclusions were often at risk of being false positive or false negative.

Implication for practice

The pre-specified effect size and α and β values for each individual meta-analysis have to be taken in the context of the conditions' seriousness, adverse effects, alternative treatments, and costs. If the required information size is not reached and the monitoring boundary not crossed, the additional number of participants required can be estimated (example box). We also propose to use a pre-specified estimate of the heterogeneity in the TSA calculation.[11] Underestimating the magnitude of heterogeneity will yield an unrealistically small information size and consequently fail to reduce the risk of spurious *P*-values. Thus, we suggest that an estimate of heterogeneity should be conservative and reflect either a moderate or large magnitude of heterogeneity.

Example box

A meta-analysis with four trials including 414 infants found that restricted compared with liberal water intake reduced mortality (relative risk reduction 0.48, 95% CI, 0.04–0.72; $P = 0.04$).[26] Applying TSA (adjusted for three previous looks, using an estimated effect size of 25% risk reduction, and heterogeneity of $I^2 = 40\%$) showed, however, an inconclusive result, i.e. the meta-analysis did not reach the estimated heterogeneity-adjusted information size ($n = 3964$) and the *Z*-curve did not cross the trial sequential monitoring boundary. The estimated number of infants required to obtain firm evidence is the difference between the accrued number of infants and the information size (equaling a difference of $n = 3550$). However, the number may be fewer if the reduction of mortality is 48% in future trials. Then the monitoring boundary will be crossed before reaching the estimated information size.



TSA does not deal with errors introduced due to the inclusion of flawed trials or biased outcome reporting. Such potential shortcomings should be appropriately assessed by subgroup, funnel-plot, and meta-regression analyses.[2,52–57] Early false results can be misleading for clinical practice regardless of the direction of the effect.[64] Focusing on harm, the TSA used should be adapted to the context. In case an intervention is in widespread use, equally conservative TSAs should be considered in assessing harm.[65] However, less conservative boundaries may be used for assessing harm when the intervention has not yet been disseminated.[65] More research on these issues are needed.

Conclusions

The interpretation of meta-analyses is complex. Many meta-analyses used as the basis for recommending interventions for newborn infants did not meet the standards for being conclusive if they had been analyzed as a single trial with interim looks. The meta-analyses needed more participants to obtain evidence for a 25% or 15% relative risk reduction. To obtain a more comprehensive assessment of meta-analyses we suggest conducting TSA that controls the risk of random error. In this way, authors and readers of meta-analyses may reach a more balanced conclusion on the effect of interventions.

Supplementary data

A better quality version of figure 3 is available at *IJE* Online.

Acknowledgements

We thank Kate Whitfield and Dimitrinka Nikolova for valuable comments and suggestions on this manuscript.

KEY MESSAGES

- Little attention is being paid to random error risk in meta-analysis.
- Trial sequential analysis is a statistical approach that may reduce random error risk in cumulative meta-analyses.
- Applying trial sequential analysis on apparently conclusive neonatal meta-analyses showed that many meta-analyses become inconclusive when adjusted for random error risk.

References

[1] Young C, Horton R. Putting clinical trials into context. *Lancet* 2005;**366:**107–8.

[2] Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions 4.2.5. Available at: http://www.cochrane.org/resources/handbook/hbook.htm (Accessed February 31, 2008).

[3] Dickersin K, Rennie D. Registering clinical trials. *JAMA* 2003;**290:**516–23.

[4] Lau J, Schmid CH, Chalmers TC. Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. *J Clin Epidemiol* 1995;**48:**45–57.

[5] Berkey CS, Mosteller F, Lau J, Antman EM. Uncertainty of the time of first significance in random effects cumulative meta-analysis. *Control Clin Trials* 1996;**17:**357–71.

[6] Pogue J, Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. *Control Clin Trials* 1997;**18:** 580–93.

[7] Pogue J, Yusuf S. Overcoming the limitations of current meta-analysis of randomised controlled trials. *Lancet* 1998;**351:**47–52.

[8] Devereaux PJ, Beattie WS, Choi PT, *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. *Br Med J* 2005;**331:** 313–21.

[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] Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983;**70:**659–63.

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

[12] Brok J, Greisen G, Jacobsen T, Gluud LL, Gluud C. Agreement between Cochrane Neonatal Group reviews and clinical guidelines for newborns at a Copenhagen University Hospital – a cross-sectional study. *Acta Paediatrica* 2007;**96:**39–43.

[13] The Cochrane Library, Issue 4, 2004. Chichester: Wiley.

[14] Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;**21:**1539–58.

[15] Fedorov V, Jones B. The design of multicentre trials. *Stat Methods Med Res* 2005;**14:**205–48.

[16] Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983;**70:**659–63.

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

[18] DeMets DL. Methods for combining randomized clinical trials: strengths and limitations. *Stat Med* 1987; **6:**341–50.

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

[20] Review Manager (RevMan) [Computer program]. Version 4.2 for Windows. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2003.

[21] Puckett RM, Offringa M. Prophylactic vitamin K for vitamin K deficiency bleeding in neonates. *Cochrane Database Syst Rev* 2000;DOI: 10.1002/14651858.CD002776.

[22] Soll RF. Prophylactic synthetic surfactant for preventing morbidity and mortality in preterm infants. *Cochrane Database Syst Rev* 1998;DOI: 10.1002/14651858.CD001079.

[23] Osborn DA, Sinn J. Formulas containing hydrolysed protein for prevention of allergy and food intolerance in infants. *Cochrane Database Syst Rev* 2004;DOI: 10.1002/14651858.CD003664.pub2.

[24] Fowlie PW, Davis PG. Prophylactic intravenous indomethacin for preventing mortality and morbidity in preterm infants. *Cochrane Database Syst Rev* 2002;DOI: 10.1002/14651858.CD000174.

[25] Yost CC, Soll RF. Early versus delayed selective surfactant treatment for neonatal respiratory distress syndrome. *Cochrane Database Syst Rev* 1999; DOI: 10.1002/14651858.CD001456.

[26] Bell EF, Acarregui MJ. Restricted versus liberal water intake for preventing morbidity and mortality in preterm infants. *Cochrane Database Syst Rev* 2001; DOI: 10.1002/14651858.CD000503.

[27] Soll RF. Multiple versus single dose natural surfactant extract for severe neonatal respiratory distress syndrome. *Cochrane Database Syst Rev* 1999; DOI: 10.1002/14651858.CD000141.

[28] Henderson-Smart DJ, Wilkinson A, Raynes-Greenow CH. Mechanical ventilation for newborn infants with respiratory failure due to pulmonary disease. *Cochrane Database Syst Rev* 2002;DOI: 10.1002/14651858.CD002770.

[29] Ho JJ, Subramaniam P, Henderson-Smart DJ, Davis PG. Continuous distending pressure for respiratory distress syndrome in preterm infants. *Cochrane Database Syst Rev* 2002; DOI: 10.1002/14651858.CD002271.

[30] Steer PA, Henderson-Smart DJ. Caffeine versus theophylline for apnea in preterm infants. *Cochrane Database Syst Rev* 1998;DOI: 10.1002/14651858.CD000273.

[31] Barrington KJ. Umbilical artery catheters in the newborn: effects of position of the catheter tip. *Cochrane Database Syst Rev* 1999; DOI: 10.1002/14651858.CD000505.

[32] Barrington KJ. Umbilical artery catheters in the newborn: effects of catheter design (end vs side hole). *Cochrane Database Syst Rev* 1999; DOI: 10.1002/14651858.CD000508.

[33] Henderson-Smart DJ, Subramaniam P, Davis PG. Continuous positive airway pressure versus theophylline for apnea in preterm infants *Cochrane Database Syst Rev* 2001;DOI: 10.1002/14651858.CD001072.

[34] Elbourne D, Field D, Mugford M. Extracorporeal membrane oxygenation for severe respiratory failure in newborn infants. *Cochrane Database Syst Rev* 2002; DOI: 10.1002/14651858.CD001340.

[35] Finer NN, Barrington KJ. Nitric oxide for respiratory failure in infants born at or near term. *Cochrane Database Syst Rev* 2001; DOI: 10.1002/14651858.CD000399.

[36] Soll RF. Prophylactic natural surfactant extract for preventing morbidity and mortality in preterm infants. *Cochrane Database Syst Rev* 1997; DOI: 10.1002/14651858.CD000511.

[37] Askie LM, Henderson-Smart DJ. Restricted versus liberal oxygen exposure for preventing morbidity and mortality in preterm or low birth weight infants. *Cochrane Database Syst Rev* 2001; DOI: 10.1002/14651858.CD001077.

[38] Soll RF, Blanco F. Natural surfactant extract versus synthetic surfactant for neonatal respiratory distress syndrome. *Cochrane Database Syst Rev* 2001; DOI: 10.1002/14651858.CD000144.

[39] Soll RF. Synthetic surfactant for respiratory distress syndrome in preterm infants. *Cochrane Database Syst Rev* 1998; DOI: 10.1002/14651858.CD001149.

[40] Darlow BA, Graham PJ. Vitamin A supplementation for preventing morbidity and mortality in very low birthweight infants. *Cochrane Database Syst Rev* 2002; DOI: 10.1002/14651858.CD000501.

[41] Andersen CC, Phelps DL. Peripheral retinal ablation for threshold retinopathy of prematurity in preterm infants. *Cochrane Database Syst Rev* 1999; DOI: 10.1002/14651858.CD001693.

[42] Ioannidis JP. Why most published research findings are false. *PloS Med* 2005;**2**(8):e124.

[43] Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. *PLoS Med* 2007;**4:**e78.

[44] WHO. Research for health – A Position paper on WHO's Role and Responsibilities in Health Research. 2006. http://www.health-policy-systems.com/content/4/1/10

[45] Egger M, Davey-Smith G, Sterne JAC. Meta-analysis. Is moving the goal post the answer? *Lancet* 1998;**351:**1517.

[46] Doust J, Del Mar C. Why do doctors use treatments that do not work? *Br Med J* 2004;**328:**474–75.

[47] McGettigan P, Henry D. Cardiovascular risk and inhibition of cyclooxygenase: a systematic review of the observational studies of selective and nonselective inhibitors of cyclooxygenase 2. *JAMA* 2006;**296:**1633–44.

[48] Jespersen CM, Als-Nielsen B, Damgaard M, *et al*. Randomised placebo controlled multicentre trial to assess short term clarithromycin for patients with stable coronary heart disease: CLARICOR trial. *Br Med J* 2006;**332:**22–27.

[49] Bjelakovic G, Nikolova D, Gluud LL, Simonetti RG, Gluud C. Mortality in randomized trials on antioxidants supplements for primary and secondary prevention. *JAMA* 2007;**297:**842–57.

[50] Gluud C, Brok J, Gong Y, Koretz R. Hepatology may have problems with putative surrogate outcome measures. *J Hepatol* 2007;**46:**734–42.

[51] Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta analyses. *Br Med J* 2007;**335:**914–6.

[52] Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. *Health Technol Assess* 2000;**4:**1–125.

[53] Schulz KF, Chalmers I, Hayes, R, Altman DG. Empirical evidence of bias. Dimensions of methodological quality

associated with estimates of treatment in controlled trials. *JAMA* 1995;**273:**408–12.

[54] Moher D, Pham B, Jones A *et al*. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses. *Lancet* 1998;**352:** 609–13.

[55] Kjaergard LL, Villumsen J, Gluud C. Reported methodological quality and discrepancies between large and small randomized trials in meta-analyses. *Ann Intern Med* 2001;**135:**982–9.

[56] Als-Nielsen B, Gluud LL, Gluud C. Methodological quality and treatment effects in randomised trials - a review of six empirical studies [abstract]. 12th International *Cochrane Colloquium*, Ottawa 2004. Available at: http://www.cochrane.org/colloquia/abstracts/ottawa/O-072.htm (Accessed February 31, 2008).

[57] Wood L, Egger M, Gluud LL *et al*. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ* 2008;**336:**601–5.

[58] Moja LP, Telaro E, D'Amico R, Moshetti I, Coe L, Liberati A. Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study. *Br Med J* 2005;**330:** 1053.

[59] Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004;**291:**2457–65.

[60] Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Methods in health service research. An introduction to bayesian methods in health technology assessment. *Br Med J* 1999;**319:**508–12.

[61] Lan KK, Hu M, Cappelleri JCC. Applying the law of iterated logarithm to cumulative meta-analysis of continues endpoint. *Statistica Sinica* 2003;**13:**1135–45.

[62] Hu M, Cappelleri JCC, Lan KK. Applying the law of iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes. *Clinical Trials* 2007; **4:**329–340.

[63] 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;**64:**763–9.

[64] Montori VM, Devereaux PJ, Adhikari NK *et al*. Randomised trials stopped early for benefit: a systematic review. *JAMA* 2005;**294:**2203–9.

[65] Pocock S. Current controversies in data monitoring for clinical trials. *Clinical Trials* 2006;**3:**513–21.

Published by Oxford University Press on behalf of the International Epidemiological Association

International Journal of Epidemiology 2009;**38**:298–303
doi:10.1093/ije/dyn265

Commentary: Which meta-analyses are conclusive?

Eveline Nüesch[1,2] and Peter Jüni[1,2]*

Accepted 11 November 2008

In 1991, a meta-analysis of seven small-scale trials of intravenous magnesium in a total of 1266 patients with suspected acute myocardial infarction indicated a >50% reduction in the risk of death associated with magnesium (relative risk 0.48, 95% CI 0.26–0.88).[1] Yusuf *et al*. updated this meta-analysis in 1993[2] to include LIMIT-2,[3] at the time the only adequately sized trial, with a power of 80% to detect a moderate to large relative reduction in the risk of death of 33% associated with magnesium. Based on a total of eight trials in 3617 patients with a pooled relative risk of 0.59 (95% CI 0.38–0.91), the authors concluded that 'intravenous magnesium is a safe, effective, widely practicable and inexpensive intervention that has the potential of making an important impact on the management of patients with myocardial infarction'.[2] In 1995, ISIS-4 became available,[4] a large-scale trial in 58 050 patients, which had nearly 95% power to detect a small, but potentially clinically relevant reduction in the relative risk of death of 10% associated with magnesium. ISIS-4 clearly refuted the earlier meta-analyses and showed a trend towards more deaths in the patients allocated to magnesium, with the lower limit of the 95% CI excluding any relevant benefit of the intervention (relative risk 1.05, 95% CI 0.99–1.12).

[1] Institute of Social and Preventive Medicine, University of Bern, Switzerland.
[2] CTU Bern, Bern University Hospital, Switzerland.
* Corresponding author. Institute of Social and Preventive Medicine, University of Bern, Switzerland.
E-mail: juni@ispm.unibe.ch