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

The case of magnesium in acute myocardial infarction cast serious doubts on the trustworthiness of meta-analyses. Which meta-analyses were conclusive and which were likely to be refuted by subsequent large-scale trials? Intrigued by the magnesium example, Egger and Davey Smith[5] suggested in 1995 that funnel plots could have been used as a diagnostic tool, in which estimates of treatment effect obtained in trials included in the magnesium meta-analyses[1,2] were plotted against a measure of sample size or statistical precision, to detect bias associated with small trials. In the absence of bias, the plot will typically resemble a symmetrical inverted funnel with the results of smaller trials more widely scattered than those of larger, more precise trials. Publication bias,[6] and poor design, execution and analysis of small trials[7] may result in skewed funnel plots. Visual inspection of the funnel plot of magnesium trials and a formal statistical test of its asymmetry indicated that the funnel plot was clearly asymmetrical before ISIS-4 became available.[5,7]

In 1997, Pogue and Yusuf[8,9] took a different approach and suggested that multiple looks in meta-analyses of randomized trials may be interpreted similarly to interim looks in a single trial. The problem of interim looks in a single trial was originally addressed by Armitage[10] and Pocock[11] by group sequential analysis. Lan and DeMets[12] extended the suggested concept with an alpha-spending function to allow flexible unplanned monitoring in a trial. They introduced the cumulative *z*-curve modelled as a Brownian motion and an alpha-spending function according to O'Brien and Flemming[13] for the construction of monitoring boundaries. If a treatment effect larger than expected occurs, a trial should be terminated early when the cumulative *z*-curve for this treatment effect crossed the constructed sequential monitoring boundary. In early stages of a trial when data are sparse, only very extreme results corresponding to extreme *z*-values are accepted to indicate premature termination of a trial. The monitoring boundaries become less stringent as more data accumulate and the planned sample size of the trial is approached. The same principle could be applied to meta-analyses to determine when a meta-analysis is conclusive. Only extreme results leading to *z*-values that cross highly stringent boundaries should be accepted if little information was accrued in a meta-analysis of few, small-scale trials. Boundaries should become less stringent as more information accumulates.[8,9] In a cumulative meta-analysis of 10 magnesium trials, Pogue and Yusuf[8] found that the cumulative *z*-curve of the meta-analysis did not cross the specified monitoring boundary for overall mortality before ISIS-4[4] and suggested that the meta-analysis was not conclusive. However, Egger *et al*. identified 15 trials of magnesium in myocardial infarction published before ISIS-4.[4] When based on all 15 trials, rather than the 10 trials selected by Pogue and Yusuf, the meta-analysis crossed the monitoring boundary and became conclusive, although the results were still contradicted by ISIS-4.[14] Pogue and Yusuf's approach failed to become widely adopted.

Recently, Wetterslev *et al*. coined the term 'trial sequential analysis' for an extension of Pogue and Yusuf's approach, which reflects an increase in uncertainty if heterogeneity between trials is present in a meta-analysis.[15] In this issue, two articles by the same group use trial sequential analysis to determine whether results of published meta-analyses in neonatology[16] and across different fields[17] are conclusive. Accounting for the observed heterogeneity between trials, they find a substantial proportion of published meta-analyses potentially inconclusive. In both articles,[16,17] the authors point out that trial sequential analysis does not deal with systematic errors resulting from the inclusion of flawed trials[18] and outcome reporting[19] or publication biases[20] and that these sources of systematic errors should be appropriately examined using funnel plots[21] and analyses stratified according to methodological characteristics of trials accompanied by appropriate tests for interaction between trial characteristic and effect estimates.[22]

Here, we re-analyse the trials of intravenous magnesium in acute myocardial infarction to determine how the different diagnostic measures—funnel plots, stratified analyses according to methodological characteristics of trials and heterogeneity-adjusted trial sequential analysis—contribute to our understanding of bias and inconclusive results at four stages of the meta-analysis: (A) trials available until 1991, before LIMIT-2;[3] (B) trials until 1995, before ISIS-4[4] became available; (C) all trials until 1995, including ISIS-4;[4] and (D) all trials available to date.[14,24] Figure 1 presents funnel plots of effect sizes on the horizontal axis against their standard errors on the vertical axis, displaying asymmetry as regression lines with 95% confidence bands derived from predicting the treatment effect from univariable meta-regression analysis with the standard error as the explanatory variable.[21] Visual inspection of funnel plot and regression line suggest asymmetry at all four stages A–D of the meta-analysis, but Egger's test for funnel plot asymmetry[23] becomes positive only at stage B, after the inclusion of LIMIT-2,[3] the only adequately sized trial at that time. In subsequent stages, the shape of the funnel plot remains essentially unchanged and Egger's test for asymmetry positive, suggesting bias.

Table 1 presents the results from corresponding stratified analyses according to concealment of allocation and sample size. At stage A, stratified analyses using a fixed- and a random-effects models indicate no relevant differences between trials with adequate concealment and the remaining trials, whereas no adequately sized trials with sample sizes of 2200 patients or more were available. At stage B, after LIMIT-2[3] became available, differences become

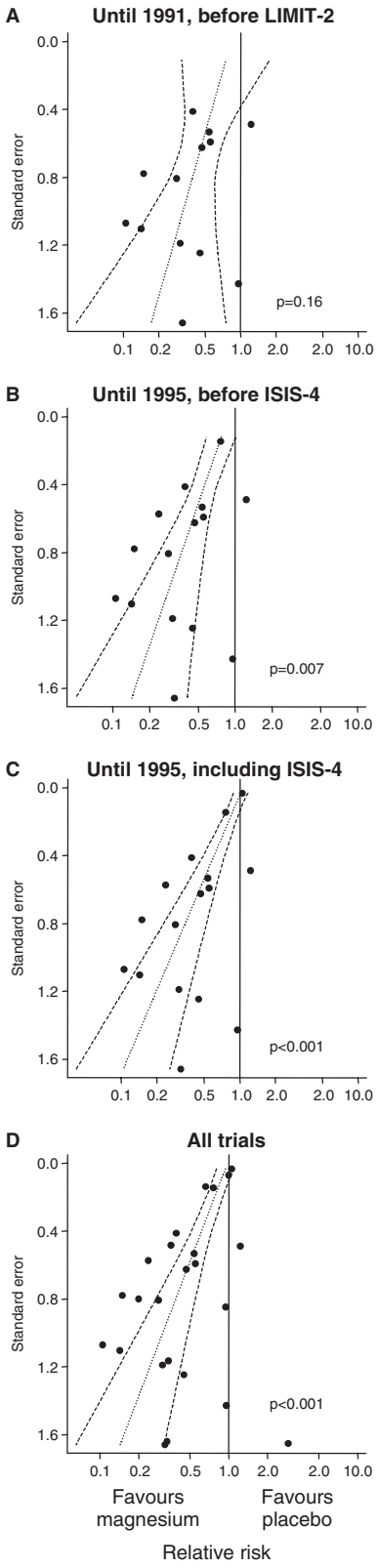

Figure 1 Funnel plots. Funnel plots are presented (**A**) for trials published until 1991, before LIMIT-2 became available; (**B**) until 1995, before ISIS-4 became available; (**C**) until 1995, including ISIS-4; and (**D**) up to 2004. Dotted lines indicate predicted treatment effects (regression line) from univariable meta-regression by using standard error as explanatory variable; dashed lines represent 95% CI. Regression lines are truncated at standard errors typically found in adequately sized trials with sufficient power to detect a moderate to large relative risk reduction of 30–40% (stages A and B) and at the standard error found in the largest trial included in the meta-analysis (stages C and D). *P*-values are derived from Egger's test for funnel plot asymmetry.[23]

apparent between trials with and without concealment of allocation and between large and small trials, but pooled effects are statistically significant in all stratified analyses and interaction tests are positive only in fixed-effect meta-analyses. With the inclusion of ISIS-4,[4] the between trial heterogeneity becomes prominent. Therefore, random-effects models attribute considerably more weight to smaller studies than fixed-effect models and results from fixed- and random-effects meta-analyses including all trials are discordant: there is still a clinically relevant mortality reduction according to the random-effects, but a clear-cut null result according to the fixed-effect meta-analysis. Even in the presence of high between-trial heterogeneity, random- and fixed-effect models show concordant results if stratified according to trial size: no effect in adequately sized trials and an unrealistically large beneficial effect of magnesium on overall mortality in small trials. Positive tests of interaction in both random- and fixed-effect analyses indicate that these differences between adequately sized and small trials are unlikely to have occurred by chance alone.

Figure 2 presents results from trial sequential analysis using fixed-effect meta-analysis (top) and random-effects meta-analysis (bottom). The dashed horizontal line represents the monitoring boundaries to be reached by the z-value of a meta-analysis to indicate that results are conclusive before the number of 24 899 patients is reached, which is necessary to detect a relative risk reduction of 15% with 80% power at a two-sided α of 0.01. The boundary becomes less stringent with more patients accruing and will converge to a z-value of 2.58 corresponding to the α-level of 0.01 indicating conclusive results when sufficient numbers of patients have been accumulated. Neither in random-effects, nor in fixed-effect meta-analyses, the z-curve crosses the boundary before ISIS-4 becomes available and the necessary information size of nearly 25 000 patients is reached, suggesting that the results of both, random- and fixed-effect meta-analyses were inconclusive. After inclusion of ISIS-4,[4] however, results are conflicting: evidence of a null effect according to the fixed-effect model, but evidence of a benefit of magnesium according to the random-effects model, which vanishes only after the analysis is restricted to trials with adequate sample size (data available on request).

It is the overall pattern found in funnel plots, stratified analyses and heterogeneity-adjusted trial sequential analysis, which provides a clear-cut insight into the trustworthiness of the different stages of the meta-analysis of magnesium in acute myocardial infarction.[1,2,14,24] At stage A, formal tests of funnel plot asymmetry and interaction tests accompanying stratified analyses are still negative due to a lack of power and some would have concluded that the evidence accumulated was unbiased and trustworthy. Heterogeneity-adjusted trial sequential analysis unequivocally indicates, however, that the evidence was inconclusive at this stage. At stage B, trial sequential analysis suggests that the accumulated evidence is still unconvincing even though LIMIT-2[3] was included. In addition, the test for funnel plot asymmetry becomes positive. At stages C and D, after the inclusion of ISIS-4,[4] heterogeneity-adjusted trial sequential analyses of

Table 1 Stratified analyses

			Random-effects meta-analysis		Fixed-effect meta-analysis		
	Number of trials	Number of patients	Relative risk (95% CI)	*P*-value for interaction	Relative risk (95% CI)	*P*-value for interaction	Heterogeneity I^2 (%)
A. Until 1991, before LIMIT-2[14]							
Overall	13	2028	0.46 (0.32–0.65)		0.42 (0.30–0.59)		0.0
Concealment of allocation				0.99		0.89	
Adequate	2	551	0.45 (0.20–1.02)		0.44 (0.20–0.98)		0.0
Inadequate or unclear	11	1477	0.45 (0.29–0.69)		0.41 (0.29–0.60)		10.7
Sample size				–		–	
⩾2200 patients	0	0	–		–		–
<2200 patients	13	2028	0.46 (0.32–0.65)		0.42 (0.30–0.59)		0.0
B. Until 1995, before ISIS-4[14]							
Overall	15	4559	0.48 (0.34–0.67)		0.57 (0.47–0.70)		30.6
Concealment of allocation				0.64		0.036	
Adequate	4	2531	0.50 (0.28–0.91)		0.67 (0.52–0.85)		50.0
Inadequate or unclear	11	2028	0.45 (0.29–0.69)		0.41 (0.29–0.60)		10.7
Sample size				0.094		0.002	
⩾2200 patients	1	2316	0.76 (0.59–0.99)		0.76 (0.59–0.99)		0.0
<2200 patients	14	2243	0.43 (0.31–0.60)		0.39 (0.29–0.54)		0.0
C. Until 1995, including ISIS-4[14]							
Overall	16	62 609	0.53 (0.38–0.75)		1.01 (0.95–1.06)		66.8
Concealment of allocation				0.25		<0.001	
Adequate	5	60 581	0.69 (0.46–1.03)		1.03 (0.97–1.09)		77.3
Inadequate or unclear	11	2028	0.45 (0.29–0.69)		0.41 (0.29–0.60)		10.7
Sample size				0.007		<0.001	
⩾2200 patients	2	60 366	0.92 (0.67–1.26)		1.04 (0.98–1.10)		82.4
<2200 patients	14	2243	0.43 (0.31–0.60)		0.39 (0.29–0.54)		0.0
D. All trials[14,24]							
Overall	24	72 920	0.65 (0.53–0.80)		0.98 (0.93–1.03)		65.8
Concealment of allocation				0.40		<0.001	
Adequate	9	67 945	0.80 (0.65–0.98)		1.02 (0.97–1.07)		71.7
Inadequate or unclear	15	4795	0.56 (0.43–0.74)		0.58 (0.47–0.71)		7.6
Sample size				0.001		<0.001	
⩾2200 patients	4	69 758	0.89 (0.75–1.06)		1.01 (0.97–1.07)		83.0
<2200 patients	20	2982	0.42 (0.32–0.57)		0.39 (0.30–0.52)		0.0

Results from stratified analysis according to allocation concealment and sample size are presented using fixed- and random-effects models including trials published until 1991 and before LIMIT-2; until 1995 and before ISIS-4, until 1995 including ISIS-4 and up to 2004. *P*-values for interaction between treatment effect and trial characteristics were derived using meta-regression for random-effects models and *z*-tests for fixed-effect models.

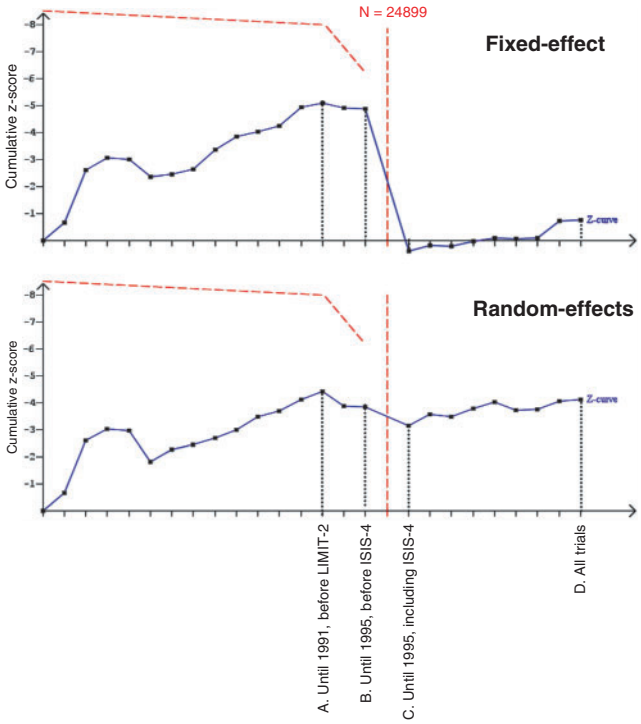

Figure 2 Heterogeneity-adjusted trial sequential analysis Trial sequential analysis of trials of intravenous magnesium using fixed-effect (top) and random-effects meta-analysis (bottom). The dashed vertical line indicates that the number of patients necessary to detect a relative risk reduction of 15% with 80% power at $\alpha = 0.01$ is 24 899 if a baseline risk of 10% and a heterogeneity between trials of $I^2 = 30\%$ are assumed. The dashed horizontal line represents the monitoring boundaries to be reached by the *z*-value of a meta-analysis to indicate that results are conclusive before the necessary number of 24 899 patients is reached. The boundary becomes less stringent when more trials and patients are included and will converge to a *z*-value of 2.58, corresponding to the α-level of 0.01, to indicate conclusive results when sufficient numbers of patients are accumulated

random- and fixed-effects meta-analyses are discordant. Here, the appropriately powered tests of funnel plot asymmetry and tests of interaction between sample size and treatment effect indicate that the inclusion of trials of inadequate size leads to a severe distortion of results.

Egger and Davey Smith concluded in 1995 that 'results of meta-analyses that are exclusively based on small trials should be distrusted - even if the combined effect is statistically highly significant. Several medium-sized trials of high quality seem necessary to render results trustworthy.'[5] These conclusions still hold in 2009. If appropriately used and interpreted, funnel plots with formal statistical tests of asymmetry, stratified analyses accompanied by tests of interaction and heterogeneity-adjusted trial sequential analyses will all contribute to our understanding about when to consider a meta-analysis conclusive.

Acknowledgements

We are grateful to Kristian Thorlund, Jørn Wetterslev and Christian Gluud for help with trial sequential analysis of the magnesium trials and for stimulating discussions.

Conflict of interest: None declared.

References

[1] Teo KK, Yusuf S, Collins R, Held PH, Peto R. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. *Br Med J* 1991;**303:**1499–503.

[2] Yusuf S, Teo K, Woods K. Intravenous magnesium in acute myocardial infarction. An effective, safe, simple, and inexpensive intervention. *Circulation* 1993;**87:**2043–46.

[3] Woods KL, Fletcher S, Roffe C, Haider Y. Intravenous magnesium sulphate in suspected acute myocardial infarction: results of the second Leicester Intravenous Magnesium Intervention Trial (LIMIT-2). *Lancet* 1992;**339:**1553–58.

[4] ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group. ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58 050 patients with suspected acute myocardial infarction. *Lancet* 1995;**345**:669–85.

[5] Egger M, Davey Smith G. Misleading meta-analysis. *Br Med J* 1995;**310:**752–54.

[6] Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991; **337:**867–72.

[7] Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;**315:**629–34.

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

[9] Pogue JM, Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. *Control Clin Trials* 1997; **18**:580–93; discussion 661–66.

[10] Armitage P. Sequential analysis in therapeutic trials. *Annu Rev Med* 1969;**20:**425–30.

[11] Pocock S. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977;**64:**191–99.

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

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

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

[15] 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.

[16] 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.

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

[18] 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. *Br Med J* 2008;**336:**601–05.

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

[20] Dwan K, Altman DG, Arnaiz JA *et al*. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE* 2008;**3:**e3081.

[21] Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol* 2001;**54:**1046–55.

[22] Jüni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *Br Med J* 2001;**323:**42–46.

[23] Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med* 2006;**25:**3443–57.

[24] Li J, Zhang Q, Zhang M, Egger M. Intravenous magnesium for acute myocardial infarction. *Cochrane Database Syst Rev* 2007:CD002755.