

## Review

# Bench-to-bedside review: Avoiding pitfalls in critical care meta-analysis – funnel plots, risk estimates, types of heterogeneity, baseline risk and the ecologic fallacy

Michael C Reade<sup>1,2</sup>, Anthony Delaney<sup>3</sup>, Michael J Bailey<sup>4</sup> and Derek C Angus<sup>1</sup>

<sup>1</sup>CRISMA Laboratory, Department of Critical Care Medicine, University of Pittsburgh, 3550 Terrace Street, Pittsburgh, PA, 15261, USA

<sup>2</sup>Current address: Austin Hospital, University of Melbourne, 145 Studley Road, Heidelberg 3084, Australia

<sup>3</sup>Royal North Shore Hospital and Northern Clinical School, University of Sydney, Pacific Highway, St Leonards, NSW, 2065, Australia

<sup>4</sup>Australian and New Zealand Intensive Care Research Centre, Department of Epidemiology and Preventive Medicine, Monash University, The Alfred Hospital, Commercial Road, Melbourne 3004, Australia

Corresponding author: Michael C Reade, [mreade@doctors.net.uk](mailto:mreade@doctors.net.uk)

Published: 25 July 2008

This article is online at <http://ccforum.com/content/12/4/220>

© 2008 BioMed Central Ltd

*Critical Care* 2008, **12**:220 (doi:10.1186/cc6941)

## Abstract

Meta-analysis can be a powerful tool for demonstrating the applicability of a concept beyond the context of individual clinical trials and observational studies, including exploration of effects across different subgroups. Meta-analysis avoids Simpson's paradox, in which a consistent effect in constituent trials is reversed when results are simply pooled. Meta-analysis in critical care medicine is made more complicated, however, by the heterogeneous nature of critically ill patients and the contexts within which they are treated. Failure to properly adjust for this heterogeneity risks missing important subgroup effects in, for example, the interaction of treatment with varying levels of baseline risk. When subgroups are defined by characteristics that vary within constituent trials (such as age) rather than features constant within each trial (such as drug dose), there is the additional risk of incorrect conclusions due to the ecological fallacy. The present review explains these problems and the strategies by which they are overcome.

## Introduction

Meta-analysis is a tool for quantitative systematic review of observational studies and controlled trials that weights available evidence based on the numbers of patients included, the effect size, and often statistical tests of agreement with other trials. Meta-analysis may be particularly suited to critical care medicine. Trials in intensive care typically enrol patients with a variety of pathologies, which can make demonstrating treatment efficacy difficult. These trials are usually underpowered for subgroup analyses. Multicentre trials can increase power with more patients, but between-centre heterogeneity can limit this benefit. Although between-centre heterogeneity can be accounted for, statistical techniques are evolving and imperfect [1]. Conducting a trial in a single centre removes between-centre heterogeneity, but when such trials (for example, those of

early goal-directed therapy for severe sepsis [2] and of tight glycaemic control in critically ill patients [3]) find treatment effects, physicians can be reluctant to implement the findings if they suspect they were unique to the study institution [4,5]. The ability to quantitatively detect subgroup effects within heterogeneous populations and to demonstrate external validity should make meta-analyses fundamental components of the critical care literature.

Unfortunately, meta-analysis in critical care can be misleading. A 1998 meta-analysis found albumin use in critically ill patients associated with a 6% increase in absolute mortality [6]. A 6,997-patient randomised controlled trial could not confirm this finding [7]. Meta-analyses do not always agree, but even high-quality reviews attempting to reconcile their differences – such as the review that demonstrated the superiority of sucralfate over histamine receptor-2 antagonists [8] – have been contradicted later by definitive clinical trials [9].

Patients in critical care trials are often studied solely because of their presence in an intensive care unit, sometimes even before their diagnosis is known. Interventions can be delivered with differing fidelity and have effects dependent on the baseline risk, which is seldom constant across trials [10]. Other aspects of what constitutes intensive care are frequently highly variable. If not appropriately addressed, heterogeneity in patients, interventions and context can produce misleading conclusions. The present review highlights how these conclusions arise, and explores approaches to these problems.

## Identifying publication bias: inspecting the funnel plot

Conducting a meta-analysis using published study results has the advantage of using evidence already subjected to peer

scrutiny. Positive studies have a greater chance of being published, however, which can falsely exaggerate the results [11]. To combat publication bias, meta-analyses may also include unpublished data [12], a process facilitated by the modern requirement to prospectively register clinical trials.

Regardless of whether they incorporate unpublished data, all meta-analyses should check for inclusion bias. A funnel plot (Figure 1) graphs each study's estimated treatment effect against an estimate of the precision of this estimate [13], such as the standard error or the number of patients included. Less precise estimates should become increasingly spread out, forming a funnel (Figure 1a). Assymetry suggests omission of some studies, although study heterogeneity and the small study effect [14] can produce the same pattern (Figure 1b,c). Publication bias can also produce a hollow plot, where studies showing minimal effect are missing while strongly positive or negative studies are included (Figure 1e). Publication bias can be statistically tested and even corrected [14], but the suggested methods have not gained widespread acceptance [15].

Publication bias may be a particular problem in critical care medicine. Only 49 out of 139 critical care meta-analyses adequately reported an absence of publication bias [16]. Inspecting the funnel plot, however, is effective only when there are enough trials [13]. In critical care, with comparatively few trials for any clinical question but with a relatively open research community, the certainty that all appropriate trials have been included is at least as useful.

### Combining different measures of treatment effect: risk difference, risk ratio or odds ratio?

A drug that reduces mortality by 20% will save a greater number of people when used in a high-risk population versus a low-risk population. Displaying the range of measures of absolute effect (usually the risk difference; Table 1) in a meta-analysis demonstrates this fact, making the data intuitively accessible. If attempting to combine estimates of treatment effect in different populations, however, it is better to use measures of relative risk – such as odds ratios, risk ratios or hazard ratios [17,18].

Unfortunately the choice of a suitable summary statistic is more complicated than would appear. Odds ratios and risk ratios both reflect relative risk, but although related they are fundamentally different (Table 1). The odds ratio is the only valid measure of association in case-control studies. As meta-analysis developed from the statistical approaches used to combine case-control studies, the odds ratio has become the default measure of effect [17]. Although theoretically the mathematically superior approach, in certain circumstances there are empirical reasons for challenging this default. For example, a review of 551 meta-analyses found median study heterogeneity was lower when a pooled relative risk (rather

than odds ratio) was used [17]. In individual studies, however, many analyses were clearly better performed with one or the other, while some analyses showed no difference. The factors influencing which summary statistic was preferable could not be predicted, and neither statistic was found clearly superior overall.

### Meta-analysis avoids Simpson's paradox

It might be tempting to simply pool data from all patients in a number of seemingly similar trials as if they came from one large study. Unfortunately, this can lead to Simpson's paradox [19], where (for example) a beneficial effect in each study can become an apparently detrimental effect when the data are aggregated. In practice, Simpson's paradox only arises when pooling the results of observational studies in which there is a severe imbalance in an important confounding factor along with unequal group assignments. Randomisation in controlled trials should prevent this occurrence.

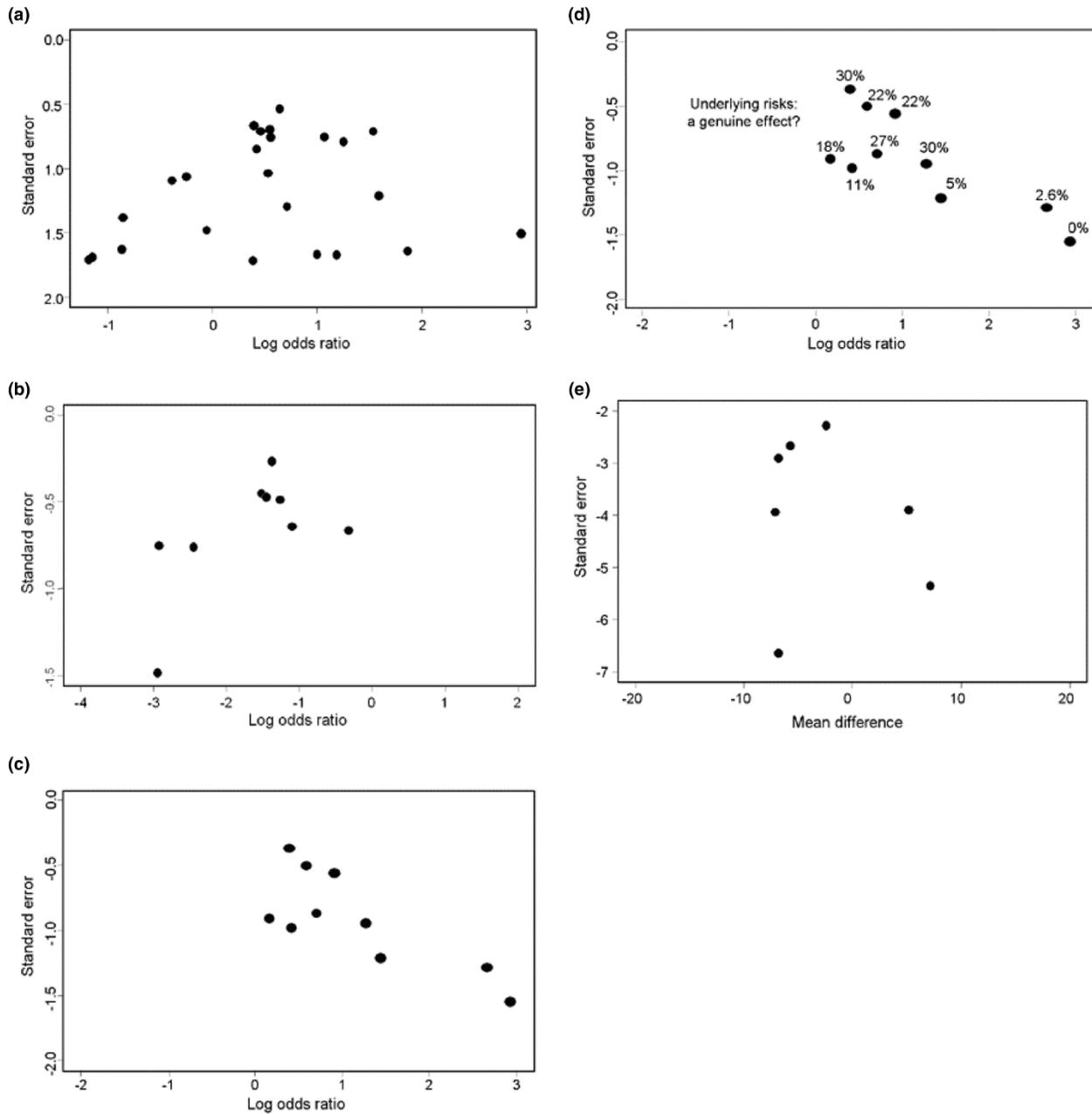
Meta-analysis overcomes Simpson's paradox by accounting for the enrolment of patients in different studies. There is a published example [20], but the three theoretical studies presented in Table 2 may be clearer. In this case, all three studies find treatment associated with lower mortality than the placebo. If the data are simply pooled, the relative risk is reversed. A fixed-effects meta-analysis finds the risk ratio to be 0.898, however, which agrees with the individual study conclusions. Figure 2 shows a graphical representation of a similar effect.

### Are studies too different to combine? Detecting heterogeneity using the $Q$ test and the $I^2$ statistic

It is not surprising that a new drug tested at a moderate dose in very ill patients receiving excellent care might have a different effect when tested at a high dose, with inconsistent delivery and monitoring, in patients across a spectrum of disease severity in a hospital struggling to provide basic services. Some trials are simply too different to combine. Heterogeneity amongst patients, contexts or interventions is problematic only if the intervention's efficacy is influenced by one of these factors. In critical care, this is often the case: for example, with activated protein C, which was effective in a mixed population containing high-risk patients [21] but not in a low-risk population [22].

If the confidence intervals of the Forest plot (Figure 3) do not overlap, the true treatment effect in those studies is probably different. Heterogeneity can also be statistically tested, most commonly using the  $Q$  test. A small  $P$  value means the null hypothesis (of study homogeneity) should be rejected – and the studies should not be combined, at least not without adjustment. Unfortunately, this test's power is relatively low when there are few studies, but is prone to overdetect heterogeneity when there are many studies. Threshold  $P$  values are arbitrarily often set higher or lower than 0.05.

**Figure 1**



Funnel plots demonstrating publication bias [15]. **(a)** Publication bias is not present, so the funnel plot should be roughly symmetrical. **(b)** If the plot is not symmetrical this may indicate publication bias, but there are other possible explanations. The small outlier study may be of lesser quality, which often results in exaggerated treatment effect sizes, or it may have been performed in a particularly high-risk population where the effect is large. **(c)** Also asymmetrical, it appears that the smaller, less precise studies are all much more positive than the larger, more precise studies. This appears a good example of publication bias. **(d)** If the control event rates are added to the plot, however, the interpretation may be different. Trials with the lowest control event rates demonstrate the most positive results. The intervention may work better in lower risk patients. Alternatively this could truly represent publication bias. From the funnel plot it is impossible to know. **(e)** The funnel plot is hollow, which is possibly publication bias of the type where significant studies are more likely to be published than those showing no difference.

A more recent approach is to report the  $I^2$  statistic [23], which quantifies the percentage of total variation between studies that is due to heterogeneity rather than due to chance. A value of 0% indicates no heterogeneity, with the

scale increasing to 100%. In contrast to the Q test, the  $I^2$  statistic facilitates the comparison of meta-analyses of different sizes. There is no value of  $I^2$  that is considered too high: the original description suggested  $I^2$  values of 25%,

**Table 1**

| Odds ratio, relative risk and risk difference |             |             |             |
|---|-------------|-------------|-------------|
|   | Treatment   | Placebo     |             |
| Alive   | a           | b           | (a + b) = e |
| Dead  | c           | d           | (c + d) = f |
|   | (a + c) = g | (b + d) = h | i           |

Risk difference = risk of death with treatment – risk of death with placebo = (c / g) – (d / h). Risk ratio = risk of death with treatment / risk of death with placebo = (c / f) / (d / b). Odds ratio = odds of death with treatment / odds of death with placebo = (c / a) / (d / b). The odds ratio approximates the relative risk when the event rate (here, death) is uncommon; however, when the event rate is common the odds ratio will overestimate the relative risk.

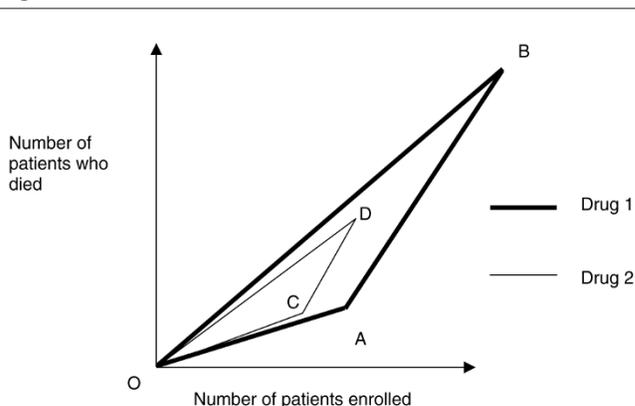
50% and 75% indicated low, moderate and high levels of heterogeneity, respectively, but that other factors (such as consistency of direction of effect and the clinical characteristics of the study) precluded definition of an arbitrary threshold.

**Dealing with heterogeneity: study selection – fixed-effects models and random-effects models**

If heterogeneity is detected, it is simplest to exclude the outlier studies, preferably with some justification. Excluding studies simply because they do not agree with the majority defeats the purpose of the meta-analysis. To counter the suspicion that inclusion criteria have been adjusted to achieve a desired effect, an analysis plan to deal with heterogeneity should be specified in advance. If heterogeneity is detected, the plan should identify which trial characteristics (such as quality, drug dose, baseline risk, and so forth) will be grounds for exclusion.

Heterogeneity is sometimes more informative than problematic. The cleanest signal would be found by looking only at trials with similar eligibility and exclusion criteria, drug doses, and hospital contexts. Showing a qualitatively consistent treatment effect despite significant heterogeneity, however, is

**Figure 2**



Demonstration of Simpson's paradox occurring if results from two clinical trials are simply pooled rather than subjected to meta-analysis. The slope of each vector represents a mortality rate (deaths / total number of patients) in patient groups taking Drug 1 (dark lines) and Drug 2 (light lines). OA and OC represent the results of one trial, and AB and CD the other trial. OA has the lesser slope, meaning Drug 1 is superior in this trial. Similarly, AB and CD demonstrate Drug 1 is superior. If the data are simply pooled, the overall effect slope (of lines OB and OD) is paradoxically reversed. The same diagram can be used to demonstrate Simpson's paradox due to subgroup effects in a single nonrandomised trial, such as that in Table 2. Here OA and OC represent results from one stratum of the trial, and AB and CD the other. When the confounding effect of stratum is ignored, Drug 2 is paradoxically superior. Adapted from [39].

equivalent to showing the treatment works in a variety of contexts – the definition of external validity.

Rather than exclude studies, the other simple approach is to statistically allow for differences between trials using a random-effects model. The intuitive assumption underlying meta-analysis is that of fixed effects: that a number of different studies are being combined to estimate one true effect of the intervention. In contrast, the random-effects model does not assume an intervention has the same effect in each of the studies. An individual study is therefore considered a random sample from a hypothetical population

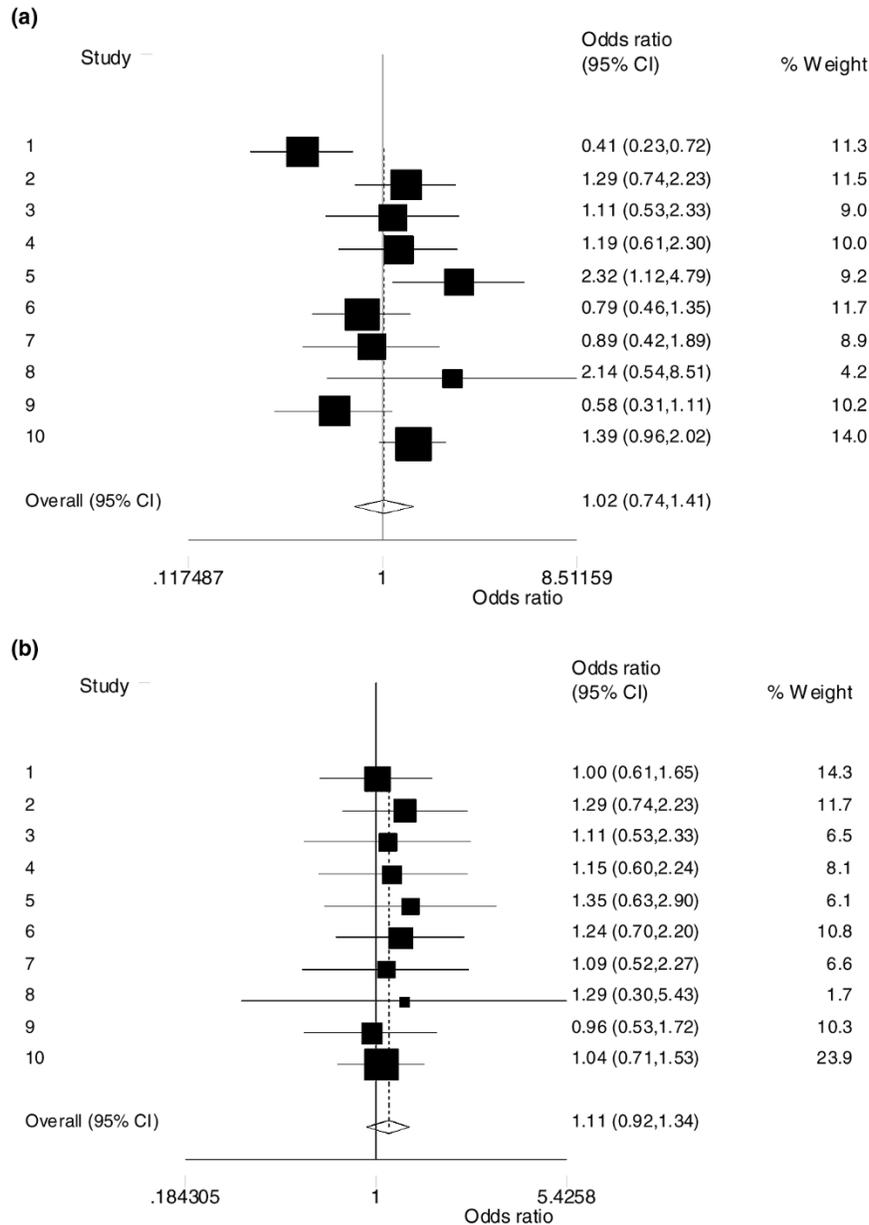
**Table 2**

|             | Mortality           |                     | Risk ratio <sup>a</sup> | Significance |
|-------------|---------------------|---------------------|-------------------------|--------------|
|             | Treatment           | Placebo             |                         |              |
| Study 1     | 97 / 492 = 19.7%    | 202 / 802 = 24.8%   | 0.794                   | 0.03         |
| Study 2     | 590 / 795 = 74.2%   | 410 / 510 = 80.4%   | 0.924                   | 0.01         |
| Study 3     | 300 / 610 = 49.2%   | 286 / 490 = 58.4%   | 0.843                   | 0.002        |
| Pooled data | 987 / 1,897 = 52.0% | 898 / 1,812 = 49.6% | 1.050                   | 0.13         |

Meta-analysis of the above data (using [40]), fixed-effects model: Q statistic = 3.398, P = 0.18; combined risk ratio = 0.898, P = 0.0001.

<sup>a</sup>Treatment / placebo.

**Figure 3**



Illustrations of Forest plots. **(a)** Statistical heterogeneity.  $I^2 = 62\%$ , heterogeneity chi-squared = 23.54 (degrees of freedom = 9),  $P = 0.005$ .

**(b)** Lack of statistical heterogeneity.  $I^2 = 0\%$ , heterogeneity chi-squared = 1.25 (degrees of freedom = 9),  $P = 0.999$ . 95% CI, 95% confidence interval. Based on Stata output using modified sample data [<http://econpapers.repec.org/software/bocbocode/s456798.htm>].

of similar studies. The study's effect estimate is consequently considered less precise.

**Informative confounding: the two types of heterogeneity**

Rather than adjust for different effects in different studies, it can be better to realise that doing so obscures potentially important information. If a drug is of benefit in men but does

some harm to women, knowing this might be more important than using a single summary statistic to conclude overall moderate advantage.

Between-study variation in the constituent trials of a meta-analysis falls into two categories: trial-level factors and patient-level factors. Trial-level factors apply to all patients in each trial: for example, drug dose. In contrast, patient-level

factors such as age and sex vary both within and between trials. Heterogeneous trial-level factors are generally easily taken into consideration, but patient-level confounding can be particularly difficult.

**Dealing with trial-level heterogeneity: subgroup meta-analysis and meta-regression**

The two approaches to dealing with trial-level heterogeneity are subgroup meta-analysis and meta-regression. Consider a number of trials of a drug, some of which use the oral route and some the intravenous route. Conducting separate meta-analyses on each of these strata makes clinical sense, particularly if there is a difference in the observed effect between the two routes. One could then conclude for each route whether the drug was beneficial. This analysis will also indirectly estimate the effect of the administration route, but the possibility of confounding (for example, perhaps patients who could use the oral route were less unwell) makes such a comparison unwise.

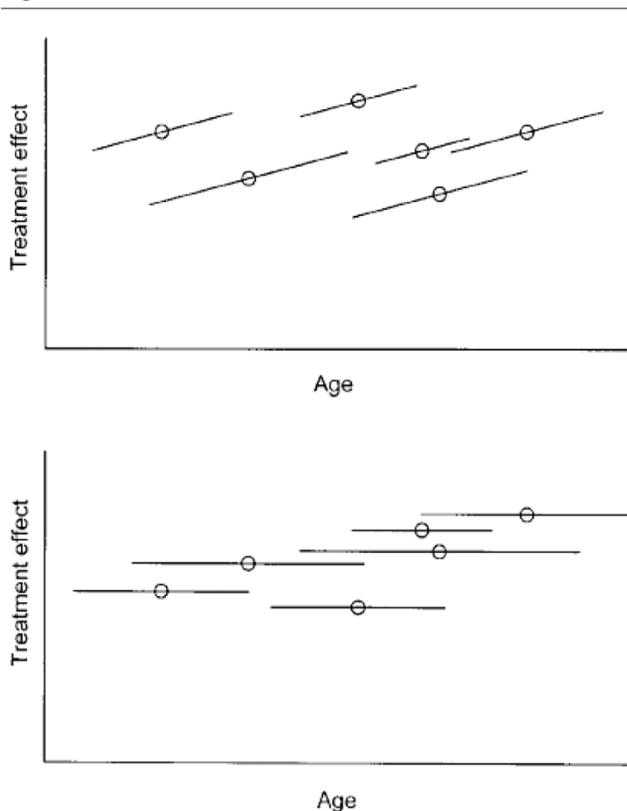
The alternative approach is meta-regression [24], a statistical model quantifying the effect of various study characteristics on the estimated overall effect. This approach is particularly useful for understanding the effect of a factor present at more than two levels, the classic example being drug dose. Meta-regression will quantify how much of the between-trial heterogeneity is explained by the various drug doses used, and is most useful when there is a significant treatment effect, a large number of studies, sufficient between-study variation in the postulated confounding variable, and sufficient heterogeneity among the treatment effects [25].

The main criticism of these techniques is that they constitute data-dredging. If trials are split into too many subgroups or too many factors are incorporated into meta-regressions, the probability of a false-positive conclusion due to multiple comparisons increases. Conversely, the small number of trials on which most of these analyses are based means power is limited, so true associations may be missed [26]. As with subgroup analyses in clinical trials, splitting a meta-analysis into subgroups is considered by many to be hypothesis generating at best. Nonetheless, the hypotheses may be stronger if the subgroups are based on prerandomisation characteristics, were planned *a priori* and allowed an adequately powered analysis, and if there is statistical adjustment for multiple testing [27].

**Dealing with patient-level heterogeneity: avoiding the ecological fallacy**

Using meta-regression to account for differences in the types of patients enrolled is possible, but potentially problematic. Such an analysis must use average patient characteristics. The relationship between the effect estimate and average patient characteristics across trials may not be the same as that relationship within trials, as is demonstrated in Figure 4. In the upper part of the figure, treatment effect is related to age within each trial, but is not related to the mean age

Figure 4



Ecological fallacy in meta-regression. Hypothetical relationships between age and treatment effect both within trials (represented by lines) and between trials (represented by dots). Upper: treatment effect is related to age within each trial, but is not related to the mean age across trials. Lower: relationship occurs across trials, but not within trials. *Stat Med*, How should meta-regression analyses be undertaken and interpreted?, Thompson SG, Higgins JP, Copyright © 2002 John Wiley & Sons Limited. Reproduced with permission [24].

across trials. In the lower part of the figure, the opposite is true: there is a relationship across trials, but not within trials. This is a classic example of the ecological fallacy, in which incorrect inferences about individual characteristics are made based upon aggregate statistics.

The term ecological fallacy was coined as an explanation for a phenomenon observed in the 1930 US census [28]. Literacy had been positively correlated with immigrant numbers in each US state, which led to the unlikely conclusion that immigrants were more literate. When literacy within each state was examined, however, the opposite relationship was observed. The explanation was that immigrants tended to settle in states where the native population was more literate.

The only way to avoid the ecological fallacy when considering possible patient-level confounding in a meta-analysis is to examine data from individuals.

## Baseline risk: an example of patient-level heterogeneity particularly relevant to critical care

It is simplest to assume the treatment in question has the same *relative* effect in patient groups with different baseline risks. This assumption can be inappropriate, particularly in intensive care, where treatments (such as activated protein C) often have substantial potential for harm as well as for benefit. If a treatment effect is heavily influenced by baseline risk (for example, harmful in low-risk patients and beneficial in those at high risk), it is necessary to adjust for risk in a multivariable model or stratified analysis. Failure to do so gives the appearance of random variation (or the variation might be misassigned to another factor, such as trial quality), whereas in reality it is an important finding. The traditional approach has been to use the event rate in the control group as a surrogate for baseline risk. This approach introduces bias due to regression to the mean, and is now considered inappropriate [29]. Alternatives have been proposed [30] but, as with other patient-level factors, the ideal solution is to investigate the interaction of treatment effect with individual patient characteristics [29], ideally using individual patient data.

## Individual patient data meta-analysis

Analysing data at an individual patient level is the most powerful meta-analytic technique available. Statistical significance is crudely determined by a ratio of explained variation over unexplained variation. The ability to account for individual patient covariates, for treatment differences between studies, and for the interactions of these factors means a greater proportion of the unexplained variation can be accounted for – increasing the power of the meta-analysis.

Other advantages of individual patient data meta-analyses over those analyses using aggregate patient data include the ability to undertake sufficiently powered exploratory subgroup analyses whilst avoiding the ecologic fallacy, to adjust for differences in baseline risk, to analyse time to event data rather than single-point outcome statistics, to update survival information, to carry out a detailed check of the primary data, and to reanalyse the data using potentially more appropriate methods [31]. Examples of such re-analyses include the ability to check the statistical assumptions of regression models, to reanalyse the data using intention-to-treat analysis, and to include patients inappropriately excluded from the original analysis.

A simulation study comparing aggregate data meta-regression and individual patient data meta-analysis found the individual patient data approach had higher statistical power. There was little agreement between the estimates of effect size between the two methods [32]. Meta-analysis of individual patient data 'is acknowledged as the gold standard' [33]. Nonetheless, individual patient data meta-analyses are performed 20 times less frequently than those using aggregate patient data [34,35], because access to detailed trial results is difficult.

## Prospective meta-analysis

Most meta-analyses are conducted retrospectively, when a series of smaller trials have failed to demonstrate a convincing result due to lack of power, or are conducted to explore subgroup effects. Where possible, the Cochrane Collaboration advocates prospective meta-analysis [36], which overcomes inconsistencies in data collection, entry criteria, study protocols and outcome measures, as well as the criticism of data-dredging. Even if individual studies are adequately powered for their primary endpoints, this is unlikely to be true for secondary outcomes and for important subgroups. Additionally, trials are conventionally funded to have 80% power, implying a 20% chance of missing a true treatment effect. Power calculations are often based on poor-quality data, and tend to be overoptimistic – such as in the recent trial of activated factor VII that hypothesised a 33% relative improvement in outcome for patients with acute intracerebral haemorrhage [37]. These factors all argue for consideration of prospective meta-analysis in the planning of any clinical trial.

## Reporting of meta-analysis

Even if appropriately conducted, a meta-analysis must be adequately reported to facilitate scrutiny of the results. Unfortunately in critical care medicine this is frequently not the case. A systematic review of 139 meta-analyses relevant to critical care found overall quality poor [16], with the most common omissions being failure to report whether a comprehensive literature search was conducted, how inclusion bias was addressed, and assessment of the validity of the included studies. The mean quality improved after publication of the Quality of Reporting of Meta-analyses (QUORUM) guidelines [38].

## Conclusion

Better awareness of the issues surrounding meta-analysis particularly relevant to critical care – especially the existence of the ecological fallacy and the possible interaction of treatment with baseline risk – will hopefully improve the performance, reporting and critical review of this valuable technique. Many pitfalls are avoided if a meta-analysis uses individual patient data and is prospectively planned, suggesting future clinical investigators should carefully consider the advantages and disadvantages of this approach.

## Competing interests

The authors declare that they have no competing interests.

This article is part of a review series on  
*Translational research*,  
edited by John Kellum.

Other articles in the series can be found online at  
[http://ccforum.com/articles/  
theme-series.asp?series=CC\\_Trans](http://ccforum.com/articles/theme-series.asp?series=CC_Trans)

## References

1. Komarek A, Lesaffre E, Legrand C: **Baseline and treatment effect heterogeneity for survival times between centers using a random effects accelerated failure time model with flexible error distribution.** *Stat Med* 2007, **26**:5457-5472.
2. Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B, Peterson E, Tomlanovich M: **Early goal-directed therapy in the treatment of severe sepsis and septic shock.** *N Engl J Med* 2001, **345**:1368-1377.
3. van den BG, Wouters P, Weekers F, Verwaest C, Bruyninckx F, Schetz M, Vlasselaers D, Ferdinande P, Lauwers P, Bouillon R: **Intensive insulin therapy in the critically ill patients.** *N Engl J Med* 2001, **345**:1359-1367.
4. Peake S, Webb S, Delaney A: **Early goal-directed therapy of septic shock: we honestly remain skeptical.** *Crit Care Med* 2007, **35**:994-995.
5. Angus DC, Abraham E: **Intensive insulin therapy in critical illness.** *Am J Respir Crit Care Med* 2005, **172**:1358-1359.
6. Bunn F, Lefebvre C, Li L, Po ALW, Roberts I, Schierhout G: **Human albumin administration in critically ill patients: systematic review of randomised controlled trials.** *Cochrane Injuries Group Albumin Reviewers.* *BMJ* 1998, **317**:235-240.
7. Finfer S, Bellomo R, Boyce N, French J, Myburgh J, Norton R: **A comparison of albumin and saline for fluid resuscitation in the intensive care unit.** *N Engl J Med* 2004, **350**:2247-2256.
8. Cook DJ, Reeve BK, Guyatt GH, Heyland DK, Griffith LE, Buckingham L, Tryba M: **Stress ulcer prophylaxis in critically ill patients. Resolving discordant meta-analyses.** *JAMA* 1996, **275**:308-314.
9. Cook D, Guyatt G, Marshall J, Leasa D, Fuller H, Hall R, Peters S, Rutledge F, Griffith L, McLellan A, Wood G, Kirby A: **A comparison of sucralfate and ranitidine for the prevention of upper gastrointestinal bleeding in patients requiring mechanical ventilation.** *Canadian Critical Care Trials Group.* *N Engl J Med* 1998, **338**:791-797.
10. Peelen L, de Keizer NF, Peek N, Bosman RJ, Scheffer GJ, de JE: **Influence of entry criteria on mortality risk and number of eligible patients in recent studies on severe sepsis.** *Crit Care Med* 2005, **33**:2178-2183.
11. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR: **Publication bias in clinical research.** *Lancet* 1991, **337**:867-872.
12. McAuley L, Pham B, Tugwell P, Moher D: **Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses?** *Lancet* 2000, **356**:1228-1231.
13. Egger M, Davey SG, Schneider M, Minder C: **Bias in meta-analysis detected by a simple, graphical test.** *BMJ* 1997, **315**:629-634.
14. Sterne JA, Egger M, Smith GD: **Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis.** *BMJ* 2001, **323**:101-105.
15. Alderson P, Green S: *Cochrane Collaboration Open Learning Material for Reviewers.* Oxford, UK: Cochrane Collaboration; 2002. [<http://www.cochrane-net.org/openlearning/index.htm>]
16. Delaney A, Bagshaw SM, Ferland A, Manns B, Laupland KB, Doig CJ: **A systematic evaluation of the quality of meta-analyses in the critical care literature.** *Crit Care* 2005, **9**:R575-R582.
17. Deeks JJ: **Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes.** *Stat Med* 2002, **21**:1575-1600.
18. Egger M, Smith GD, Phillips AN: **Meta-analysis: principles and procedures.** *BMJ* 1997, **315**:1533-1537.
19. Simpson EH: **The interpretation of interaction in contingency tables.** *J R Stat Soc* 1951, **13**:238-241.
20. Hanley JA, Theriault G: **Simpson's paradox in meta-analysis.** *Epidemiology* 2000, **11**:613-614.
21. Bernard GR, Vincent JL, Laterre PF, LaRosa SP, Dhainaut JF, Lopez-Rodriguez A, Steingrub JS, Garber GE, Helterbrand JD, Ely EW: **Efficacy and safety of recombinant human activated protein C for severe sepsis.** *N Engl J Med* 2001, **344**:699-709.
22. Abraham E, Laterre PF, Garg R, Levy H, Talwar D, Trzaskoma BL, Francois B, Guy JS, Bruckmann M, Rea-Neto A, Rossaint R, Perrotin D, Sablotzki A, Arkins N, Uttebbeck BG, Macias WL: **Drotrecogin alfa (activated) for adults with severe sepsis and a low risk of death.** *N Engl J Med* 2005, **353**:1332-1341.
23. Higgins JP, Thompson SG, Deeks JJ, Altman DG: **Measuring inconsistency in meta-analyses.** *BMJ* 2003, **327**:557-560.
24. Thompson SG, Higgins JP: **How should meta-regression analyses be undertaken and interpreted?** *Stat Med* 2002, **21**:1559-1573.
25. Schmid CH, Stark PC, Berlin JA, Landais P, Lau J: **Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors.** *J Clin Epidemiol* 2004, **57**:683-697.
26. Geddes J: **Meta-analysis in the assessment of treatment outcome.** *J Psychopharmacol* 2006, **20**:67-71.
27. Cook DJ, GebSKI VJ, Keech AC: **Subgroup analysis in clinical trials.** *Med J Aust* 2004, **180**:289-291.
28. Robinson WS: **Ecological correlations and the behavior of individuals.** *Am Sociol Rev* 1950, **15**:351-357.
29. Sharp SJ, Thompson SG, Altman DG: **The relation between treatment benefit and underlying risk in meta-analysis.** *BMJ* 1996, **313**:735-738.
30. Sharp SJ, Thompson SG: **Analysing the relationship between treatment effect and underlying risk in meta-analysis: comparison and development of approaches.** *Stat Med* 2000, **19**:3251-3274.
31. Stewart LA, Clarke MJ: **Practical methodology of meta-analyses (overviews) using updated individual patient data.** *Cochrane Working Group.* *Stat Med* 1995, **14**:2057-2079.
32. Lambert PC, Sutton AJ, Abrams KR, Jones DR: **A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis.** *J Clin Epidemiol* 2002, **55**:86-94.
33. Chalmers I: **The Cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care.** *Ann N Y Acad Sci* 1993, **703**:156-163.
34. Lyman GH, Kuderer NM: **The strengths and limitations of meta-analyses based on aggregate data.** *BMC Med Res Methodol* 2005, **5**:14.
35. Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, Thompson SG: **Meta-analysis of individual patient data from randomized trials: a review of methods used in practice.** *Clin Trials* 2005, **2**:209-217.
36. **Cochrane Collaboration** [[www.cochrane.org](http://www.cochrane.org)]
37. Mayer SA, Brun NC, Begtrup K, Broderick J, Davis S, Diringer MN, Skolnick BE, Steiner T: **Efficacy and safety of recombinant activated factor VII for acute intracerebral hemorrhage.** *N Engl J Med* 2008, **358**:2127-2137.
38. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF: **Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of reporting of meta-analyses.** *Lancet* 1999, **354**:1896-1900.
39. **Education Queensland Exploring Data** [<http://curriculum.qed.qld.gov.au/kla/eda>]
40. Hintze J: *NCSS, PASS and GESS* [computer program]. Kaysville, UT: NCSS; 2006.