Investigating cause of heterogeneity in systematic reviews

Article *in* Statistics in Medicine · June 2002

DOI: 10.1002/sim.1183 · Source: PubMed

2 authors, including:

Paul Glasziou
Bond University
701 PUBLICATIONS **34,515** CITATIONS

SEE PROFILE

STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:1503–1511 (DOI: 10.1002/sim.1183)

Investigating causes of heterogeneity in systematic reviews

P. P. Glasziou[1,*,†] and S. L. Sanders[2]

[1] *Centre for Evidence Based Health Care, University of Queensland, Brisbane, Australia*
[2] *Centre for General Practice, University of Queensland, Brisbane, Australia*

SUMMARY

What causes heterogeneity in systematic reviews of controlled trials? First, it may be an artefact of the summary measures used, of study design features such as duration of follow-up or the reliability of outcome measures. Second, it may be due to real variation in the treatment effect and hence provides the opportunity to identify factors that may modify the impact of treatment. These factors may include features of the population such as: severity of illness, age and gender; intervention factors such as dose, timing or duration of treatment; and comparator factors such as the control group treatment or the co-interventions in both groups. The ideal way to study causes of true variation is within rather than between studies. In most situations however, we will have to make do with a study level investigation and hence need to be careful about adjusting for potential confounding by artefactual factors such as study design features. Such investigation of artefactual and true causes of heterogeneity form essential steps in moving from a combined effect estimate to application to particular populations and individuals.

KEY WORDS: heterogeneity; systematic review; meta analysis; effect modification; interaction

> 'in many meta-analyses, *heterogeneity* can and should be investigated so as to increase the clinical relevance of the conclusions drawn and the scientific understanding of the studies reviewed' [1].

INTRODUCTION

Heterogeneity in systematic reviews can occur because of artefactual or real differences in treatment effect across the different studies included in the review. The real differences in effect are of great interest because of the potential to help identify who benefits most and who benefits least from a particular intervention. Berlin has presented numerous examples illustrating this point [2]. However, in order to identify real differences the potential causes of artefactual variation, such as differences in duration of follow-up or reliability of outcome measures, must first be recognized and adjusted for. Thus heterogeneity represents an

*Correspondence to: Paul Glasziou, School of Population Health, University of Queensland, Public Health Building, Herston Road, Herston, 4006, Queensland, Australia
†E-mail: p.glasziou@sph.uq.edu.au

opportunity for further understanding of treatment effects but its exploration is plagued by a number of important analytic difficulties.

We should first recognize that not everyone who undertakes an effective treatment will benefit. If, for example, an intervention has a relative risk reduction, on average, of 25 per cent, one-quarter of the expected events are averted by use of the treatment and 75 per cent of the events will still occur, despite the treatment. Three-quarters of those hoping for the benefit will not experience it. This is the case for many therapies, including aspirin, cholesterol lowering and blood pressure reduction interventions. They are clearly all useful in decreasing average risk, but they certainly do not eliminate risk.

Can we identify those people who will benefit and those people who will not? Very occasionally this is possible. For example, a single patient (*n* of 1) trial can test whether an individual patient is a responder to a particular treatment if repeated blinded cross-over of alternative treatments is used. However this is only possible in chronic diseases for which the treatment effect is transient. In the vast majority of cases we will not be able to test empirically whether an individual is a responder. Instead we must search in both systematic reviews and individual trials for factors which predict response, so called 'effect modifiers' or 'treatment interactions' (the former is the favoured term of epidemiologists, the latter of statisticians).

WHAT DO WE MEAN BY A VARYING EFFECT?

The effects of treatment may vary across different patients and settings for reasons which may be real or artefactual or both. To examine this we must first agree what is meant by 'vary'. If we say a treatment effect is constant across different underlying event rates, there are at least two broad categories of 'constant': a constant proportional effect and a constant absolute effect. Over a range of risks, we cannot simultaneously have a constant proportional and constant absolute reduction. Thus an effect may be constant on one scale but variable on the other [3]. An empirical analysis of 115 systematic reviews found that the relative risk and odds ratio appeared to be more constant than a risk difference. Schmid *et al.* [4] found that the risk difference clearly changed with baseline risk in 31 per cent of cases, whereas the relative risk and odds ratio clearly changed with baseline risk in only 13 per cent and 14 per cent of cases, respectively.

DETECTING EFFECT MODIFICATION IN SYSTEMATIC REVIEWS

We are interested in detecting factors that may produce true variation in the effects of treatments. These factors may be categorized as:

(a) the *patient* or the disease group;
(b) the *intervention* timing or intensity;
(c) the *co-intervention*, that is, what other treatments the patient is receiving;
(d) the *outcome* measurement and timing.

Unfortunately the differences between the different studies in a systematic review are not limited to these factors. Other features such as the quality of the design and conduct of

Table I. Real and artefactual causes of between-study variation in effect.

	Real	Artefactual
Patient	Disease severity Age Co-morbidity	Improper randomization Differential follow-up (non-comparable groups)
Intervention	Time Duration Dose	Non-compliance Cross-over
Co-intervention	Drugs Therapy	Undetected co-interventions
Outcome	Timing of outcome Event type	Differential and non-differential measurement error

the study, the extent of compliance with the intervention, and the accuracy of the outcome measures used may cause spurious, apparent differences in different treatment effects. These spurious differences may lead us to believe that some other factor is causing true effect modification. It is important to examine and to try to exclude such sources of difference in effect between studies before exploring the possibility of true effect modification.

Table I lists and categorizes some potential true and artefactual causes of differences in effects between different studies.

ARTEFACTUAL VARIATION

The quality of trial design has been clearly shown to affect the apparent results. Schulz *et al.* [5] have shown that improper allocation concealment is associated with odds ratios exaggerated by 41 per cent, on average. Similarly, lack of blinding of outcome assessment is associated with odds ratios exaggerated by 17 per cent, on average. These are crucial factors to consider. If, for example, the trials in one group of patients (say men) were uniformly poorly designed, and the trials in another group of patients (say women) were uniformly well designed, we may spuriously conclude that gender is a treatment effect modifier. Even if the trials have been well designed, have used proper allocation concealment, have succeeded in achieving full follow-up and blinded outcome assessment, other measurement differences may still lead to apparent differences between trials. For example, in the four large randomized trials of screening for blood in stools (faecal occult blood testing), the proportion of patients undergoing at least one screen varied between 60 per cent and 90 per cent. Clearly we cannot expect that the observed effects in such studies will be directly comparable, and methods of adjusting trials for the degree of non-compliance have been described to try to deal with this problem [6]. Of course, whether this variation is considered artefactual or not depends in part on your perspective: for the individual who will or will not comply it is artefactual, but to the policy maker non-compliance is a 'real' factor.

Variations in accuracy of the outcome measures used will also lead to differences in estimates of effects. This is illustrated in Figure 1, where a clear difference is seen in the (combined) results of trials using venography to assess the presence of deep venous thrombosis

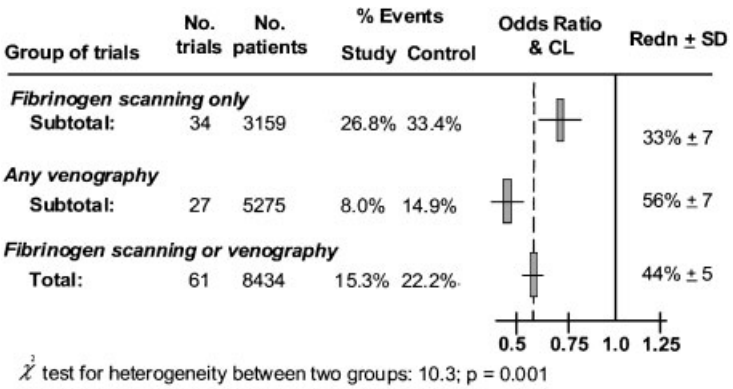

Figure 1. Artefactual effect modification. Apparent effect of antithrombotic treatment varies with measurement accuracy: fibrinogen scanning (less accurate) gives an average effect less than venography (more accurate) [7].

(DVT) and those using fibrinogen scanning [7]. Since fibrinogen scanning is less accurate than venography for detecting deep venous thrombosis, it will result in a number of false positives and false negatives, and hence dilute any real effect on true positives and true negatives. Such 'non-differential' misclassification tends to shift the estimated effect towards the null value, as is seen in Figure 1, where the effects for fibrinogen scanning are closer to a null value of 1 than the effects seen with venography.

What should we do about such artefacts? Obviously it would be ideal if the original trials could avoid these, but such universal perfection is unlikely ever to be achieved. According to Hunter and Schmidt [8], 'corrections for errors in study findings due to study imperfections (which we shall call artefacts) is essential to the development of cumulative knowledge'. Such correction is only possible, however, if we know the effects of such study imperfections, that is the degree of compliance or the sensitivity or specificity of the outcome measures. Unfortunately, for many of the biases in studies, such as poor allocation concealment, the precise effects will not be known, and hence cannot be corrected for. Subsequently, we will be left to decide whether the effects of such defects lead to sufficient bias that the study should be omitted.

TRUE EFFECT MODIFICATION

If we have satisfied ourselves that there is unlikely to be any important artefactual effect modification, we may proceed to look at our real interest – true effect modification. Except in very large trials, single studies usually have insufficient power to reliably examine effects in subgroups. By combining the results of several studies, therefore, we are more likely to have sufficient power. However, to do this reliably, the main effect should preferably have tight confidence limits: if there is insufficient power to reliably examine the main effect, there is certainly insufficient power to look for any not very large differences between subgroups. The potential causes of true effect modification are those given earlier in Table I. Let us look at a few examples from the four categories of patient, intervention, co-intervention and outcome:

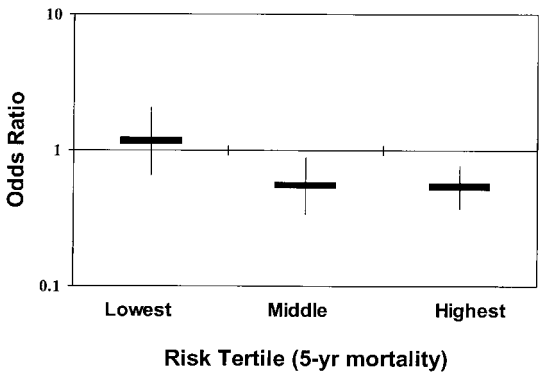

Figure 2. True effect modification by patient risk group. The odds ratio for coronary bypass versus medical treatment appears beneficial in the middle and highest tertiles but possibly harmful in the lowest tertile (p for interaction $= 0.06$) [9].

(a) *The patient*. If the intervention leads to a constant proportional reduction in effect, then the level of risk or severity of disease is important for predicting absolute effects. In this instance, high-risk patients or those with more severe disease will tend to have a greater absolute benefit. This may also be expressed by saying that a constant proportional effect implies an effect modification on the absolute scale such that those with greatest risk or severity have the greatest absolute benefit. Interaction may also occur on the proportional scale as illustrated by the analysis shown in Figure 2, where those at lowest risk appear to have no net benefit from coronary artery bypass grafting, whereas those at middle and high risk do, though the p-value for interaction here of 0.06 is not definitive [9]. This appears to be a 'qualitative' interaction with a beneficial effect at one end of the scale and a harmful effect at the other end of the scale. This has also been noted in the case of carotid endarterectomy and may be broadly true of surgical interventions, with those at very low risk not attaining sufficient benefit to outweigh the harms of surgery.

(b) *The intervention*. Many interventions will have a dose–response relationship: the effects increase with increasing intensity of the intervention (in the form of increased dose, increased duration or increased frequency). The effect will often plateau once an optimal dosage is reached and may even become harmful beyond this. Hence, exploration of such dose–response relationships is always advisable in any systematic review. It should be noted, however, that the dose–response relationship may be different for different outcomes. For example, the antiplatelet effects of aspirin are seen at minimal dosages such as 75 mg per day, the analgesic and antipyretic properties of aspirin require larger doses, and the anti-inflammatory properties larger doses still [10, 11]. As this example illustrates, there may be no single, uniform dose–response relationship for all outcomes of a single intervention. Rather, they may be outcome-specific.

The timing of therapy may also be important. A classic example is that of thrombolytic therapy for acute myocardial infarction, for which the benefits rapidly decline with time since the onset of chest pain [12]. Early institution of therapy is a reasonably obvious effect modifier, but there may be more subtle timing issues such as circadian

or menstrual cycles. For example, there is some evidence that screening mammography is more sensitive in the latter part of the menstrual cycle [13].

(c) *Co-interventions*. Other treatments may have an effect on the intervention under study. For example, the antihypertensive effects of the angiotensin-converting enzyme inhibitor enalapril vary according to whether a patient is also taking a beta-blocker [14]. This is due to a significant negative interaction for the hypotensive effect that exists when both agents are taken simultaneously.

(d) *Outcome measures and timing*. An intervention may often not have the same size of effect on all outcomes. For example, antihypertensive therapy has differential preventive effects on different types of cardiovascular disease, with the incidence of stroke being reduced more substantially than the incidence of coronary heart disease [15]. Similarly, Cox 2 inhibitors do not have a uniform effect on gastric symptoms; when compared to traditional non-steroidal anti-inflammatory drugs, they result in substantially fewer gastric ulcerations, in spite of the fact that they lead to similar rates of gastric symptoms (dyspepsia) [16].

The effect may also vary with the time of measurement of the outcome. For example, in a meta-analysis of the effects of antibiotics in acute otitis media, no effect is seen at 24 hours after starting antibiotics, but at 2–7 days, one-third fewer children still have symptoms or some degree of pain [17], see Figure 3. In this instance we are making

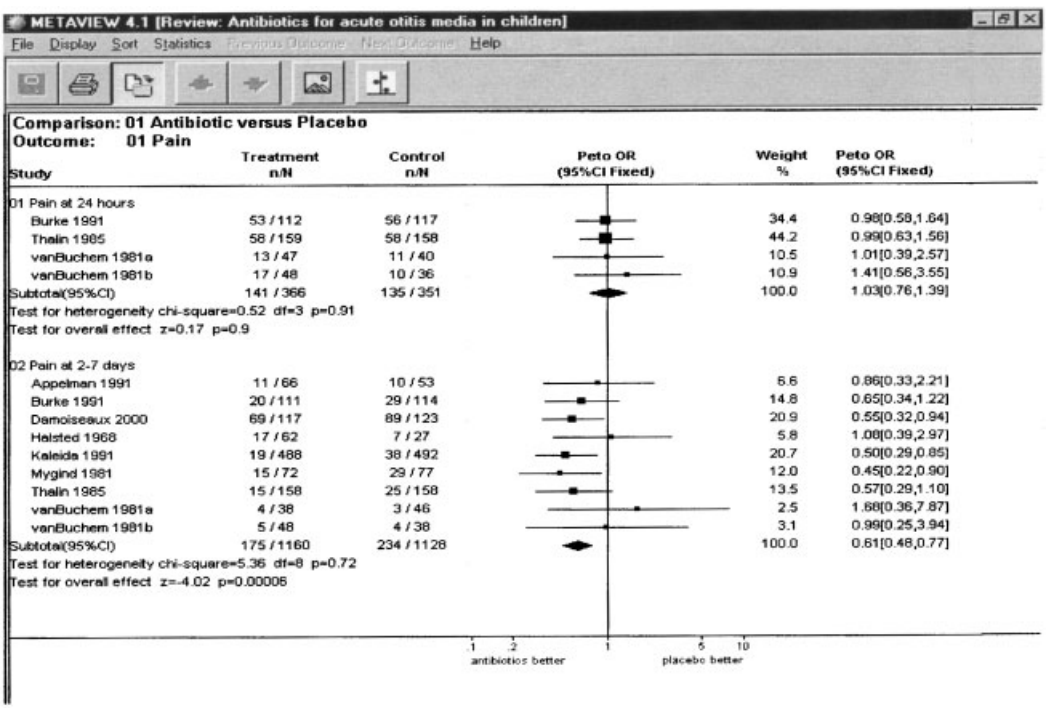

Figure 3. True effect modification by timing. The effect of antibiotics on acute otitis media is nil at 24 hours, but becomes apparent at 2–7 days [17].

comparisons between studies. Without doing an analysis within studies we cannot be certain the apparent differences are real and not the result of confounding by study design.

WITHIN-STUDY, BETWEEN-STUDY EFFECTS

How should a study of the true causes of effect modification proceed? One approach involves examining differences by study-level features – so called meta-regression [18]. While this is convenient, and often the only feasible option, it is not the most desirable because the differences between studies may be confounded by other study features.

Suppose, for example, that in the antibiotic and otitis media study referred to above, the outcomes for antibiotic A were only ascertained at 24 hours and 2 weeks, whereas for antibiotic B the outcomes were only ascertained at 3 days. If no effect of antibiotics is detected at 24 hours and 2 weeks, but an effect is detected between those time points, then the effect of different antibiotics is confounded by the study design – in this case the timing of the outcome measures. We cannot always disentangle such confounding by study design features, but it can often be reduced in systematic reviews using individual patient data.

Consider whether age might be an effect modifier in trials of antihypertensive agents for raised blood pressure. This may be either disguised or confounded by looking across the mean age of patients in different studies. Much better would be to take advantage of the within-study variation and examine the effect modification seen by age within each study. The estimate of this effect modification can then be pooled across the different studies. This issue is discussed further with examples elsewhere [19].

TRANSFERABILITY AND APPLICATION OF TRIAL RESULTS

Examining heterogeneity and effect modification is part of a broader objective of a systematic review: to decide in which, if any, future patient groups the net effect of an intervention is worthwhile. As stated earlier, even if a statistically significant overall treatment effect has been established, it is unlikely that the intervention will benefit everyone. The groups or individuals to whom the results of trials can be applied is rarely crystal clear. Common methods for assessing applicability based on the characteristics of participants in a trial, inclusion or exclusion criteria and subgroup analysis, are inadequate because they fail to account for the likely absolute benefits and harms in individuals. A better approach for identifying individuals or groups in whom the intervention is likely to produce a net benefit is based on a risk-benefit analysis examining the predictors of individual response and risk and how the risks and benefits balance [20, 21]. This analysis can be viewed as a five-step process, the first three steps relating directly to the transferability of the average treatment effect and the last two steps relating to aspects of individualizing the treatment decision:

1. *Identify the beneficial and harmful effects of the intervention*. All patient-relevant endpoints that may be influenced by the intervention should be considered, including adverse events. This is particularly important for individuals who are at low risk of

whichever adverse outcome the intervention is supposed to prevent. In such individuals, the intervention can be expected to achieve only a small absolute risk reduction, so the harms are much more important as they *may* outweigh the benefits. It is useful to tabulate all the possible positive and negative effects of the intervention, irrespective of data availability.

2. *Identify true effect modification in the relative treatment effect.* For each of the main outcomes (beneficial and harmful), we should attempt to identify any features of the population or intervention that modify the relative impact of the intervention. Ideally, this should be done using a multivariate model with appropriate interaction terms within each individual study. However, we will often have to make do with a study level meta-regression and hence need to be careful about adjusting for potential confounding by study design features. If effect modifiers are identified, then the steps below need to incorporate the effects in different groups.
3. *Assess whether the treatment effect varies with level of baseline risk.* Variation in the level of an (untreated) individual's risk of an adverse outcome is almost always an important consideration. Generally, low-risk groups will have less to gain from an intervention than high-risk groups and may not therefore experience sufficient benefit to outweigh the harms. However, examining the effect of baseline risk is subject to several statistical traps, for example, correlation between RR, OR, or RD and control group risk, regression dilution bias, and confounding by study design [22, 23].
4. *Assess the predicted absolute risk reductions for individuals.* To assess whether therapy is worthwhile, it is necessary to know the absolute magnitude of the benefit (this may be expressed as the absolute risk reduction or as the number-needed-to-treat) and harms (expressed as absolute risk increase or number-needed-to-harm). As the former is likely to vary with baseline risk it is also necessary to know the individual's expected event rate or severity. Estimates of the patient's expected event rate (PEER) can be obtained from prognostic models (for example, from cohort studies) linking values of various characteristics of the patient to the probability of the disease of interest.
5. *Weigh up the benefits and harms.* The principal issue for the individual here is whether the predicted absolute benefit has greater value than the predicted harm and cost of treatment. Consideration of the individual's preferences in relation to the potential benefits and harms is essential. If the previous step is done well, the trade-offs will often be clear. Methods developed in decision analysis, however, may also be useful. For example, quality-adjusted life-years (QALY) might provide a useful summary measure where there are trade-offs between quality and quantity of life.

CONCLUSIONS

Heterogeneity in treatment effect should be examined and explained. Exploration may reveal artefactual sources of variation resulting from poor study design, or potentially interesting 'true' sources of variation, related to characteristics of the individual, disease or intervention. Identification of true effect modifiers is an important intermediate step between combining studies and examining how the results apply to particular populations and individuals who may be at different levels of risk from those in the trial populations.

ACKNOWLEDGEMENTS

We would like to thank Iain Chalmers for helpful comments on the manuscript.

REFERENCES

1. Thompson S. In *Systematic Reviews*, Chalmers I and Altman DG (eds). BMJ Publishing Group: London, 1995; 62.
2. Berlin JA. Invited Commentary: Benefits of heterogeneity in meta-analysis of data from epidemiologic studies. *American Journal of Epidemiology* 1995; **142**(4):383–386.
3. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine* 2002; **21**:1575–1600.
4. Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Statistics in Medicine* 1998; **17**(17):1923–1942.
5. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association* 1995; **273**:408–442.
6. Glasziou PP. Meta-analysis adjusting for compliance: the example of screening for breast cancer. *Journal of Clinical Epidemiology* 1992; **45**(11):1251–1256.
7. Rodgers A, MacMahon S. Systematic underestimation of treatment effects as a result of diagnostic test inaccuracy: Implications for the interpretation and design of thromboprophylaxis trials. *Thrombosis and Haemostasis* 1995; **73**(2):167–171.
8. Hunter JE, Schmidt FL. *Methods of Meta-analysis: Correcting Error and Bias in Research Findings.* Sage: Newbury Park, 1990.
9. Yusuf S, Zucker D, Peduzzi P, Fisher, LD, Takaro T, Kennedy, JW, Davis K, Killip T, Passamani E, Norris R, Morris C, Mathur V, Varnauskas E, Chalmers TC. Effect of coronary artery bypass graft surgery on survival: overview of 10-year results from randomised trials by the Coronary Artery Bypass Graft Surgery Trialists Collaboration. *Lancet* 1994; **344**:563–570.
10. Antiplatelet Trialists' Collaboration. Collaborative overview of randomised trials of antiplatelet therapy – I: Prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients. *British Medical Journal* 1994; **308**:81–106.
11. Edwards JE, Oldman A, Smith L, Collins SL, Carroll D, Wiffen PJ, McQuay HJ, Moore RA. Single dose oral aspirin for acute pain. The Cochrane Library 2000; Issue 3, Update Software, Oxford.
12. Fibrinolytic Therapy Trialists' (FTT) Collaborative Group. Indications for fibrinolytic therapy in suspected acute myocardial infarction: collaborative overview of early mortality and major morbidity results from all randomised trials of more than 1000 patients. *Lancet* 1994; **343**:311–322.
13. Baines CJ. Menstrual cycle variation in mammographic breast density: So who cares? *Journal of the National Cancer Institute* 1998; **90**(12):875–876.
14. Wing LM, Chalmers JP, West MJ, Bune AJ, Russell AE, Elliott JM, Morris MJ. Treatment of hypertension with enalapril and hydrochlorothiazide or enalapril and atenolol: contrasts in hypotensive interactions. *Journal of Hypertension* 1987; **5**(5):S603–606.
15. Collins R, Peto R, MacMahon S, Hebert P, Fiebach NH, Eberlein KA, Godwin J, Qizilbash N, Taylor JO, Hennekens CH. Blood pressure, stroke, and coronary heart disease. Part 2, Short-term reductions in blood pressure: overview of randomised drug trials in their epidemiological context. *Lancet* 1990; **335**:827–838.
16. Silverstein FE, Faich G, Goldstein JL, Simon LS, Pincus T, Whelton A, Makuch R, Eisen G, Agrawal NM, Stenson WF, Burr AM, Zhao WW, Kent JD, Lefkewith JB, Verburg KM, Geis GS. Gastrointestinal toxicity with celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis: The CLASS Study: A randomized controlled trial. *Journal of the American Medical Association* 2000; **284**(10):1247–1255.
17. Glasziou PP, Hayem M, Del Mar CB. Antibiotics for acute otitis media in children. The Cochrane Library 2000; Issue 3. Update Software: Oxford.
18. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine* 2002; **21**:1531–1558.
19. Berlin J. *Statistics in Medicine* **21**(11):1501–1502.
20. Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. *British Medical Journal* 1995; **311**:1356–1359.
21. Lubsen J, Tijssen JG. Large trials with simple protocols: indications and contraindications. *Controlled Clinical Trials* 1989; **10**(Suppl):151S–160S.
22. Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. *British Medical Journal* 1996; **313**(7059):735–738.
23. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine* 1997; **16**(23):2741–2758.

