

USEFUL WEBSITES

- See first article in this series (*BMJ* 2006;333:83-6, doi: 10.1136/bmj.333.7558.83) for a list of general websites
- British Fertility Society (www.britishfertilitysociety.org.uk/humanfertility/index.html)—information source for health and science professionals involved in human fertility

Competing interests: None declared.

- 1 Smellie WSA, Forth J, Sundar S, Kalu E, McNulty CAM, Sherriff E, et al. Best practice in primary care pathology: review 4. *J Clin Pathol* 2006;59:893-902.
- 2 Gangar K, Cust M, Whitehead MI. Symptoms of oestrogen deficiency associated with supraphysiological plasma oestradiol concentrations in women with oestradiol implants. *BMJ* 1989;299:601-2.
- 3 Smellie WSA, Forth J, McNulty CAM, Hirschowitz L, Lilić D, Gosling R,

- et al. Best practice in primary care pathology: review 2. *J Clin Pathol* 2006;59:113-20.
- 4 National Institute for Clinical Excellence. *Fertility: assessment and treatment for people with fertility problems*. London: NICE, 2004. (Clinical guideline 11.)
- 5 Royal College of Obstetricians and Gynaecologists. *The initial investigation and management of the infertile couple*. London: RCOG Press, 1998.
- 6 Scott R, Toner J, Muasher S, Oehninger S, Robinson S, Rosenwaks Z. Follicle-stimulating hormone levels on cycle day 3 are predictive of in vitro fertilization outcome. *Fert Steril* 1989;51:651-4.
- 7 Kim YK, Wasser SK, Fugimoto VJ, Klein NA, Moore DE, Soules MR. Utility of FSH:LH ratio in predicting reproductive age in normal women. *Hum Reprod* 1997;12:1152-5.
- 8 Fox R, Conigan E, Thomas PA, Hull MG. The diagnosis of polycystic ovaries in women with oligomenorrhoea: predictive power of endocrine tests. *Clin Endocrinol* 1991;34:127-31.
- 9 Robinson S, Rodin DA, Deacon A, Wheeler MJ, Clayton RN. Which hormone tests for the diagnosis of polycystic ovarian syndrome. *Br J Obstet Gynaecol* 1992;99:232-8.

Accepted: 9 November 2006

CLINICAL EPIDEMIOLOGY NOTES

What is heterogeneity and is it important?

John Fletcher

Three simple examples from recent *BMJ* papers illustrate the importance of heterogeneity in a systematic review and how readers can assess it

Three systematic reviews published in the *BMJ*, including one in this issue, have referred to heterogeneity and dealt with it in three different ways.^{1 2 3} So what is heterogeneity, and how do we assess its importance in a systematic review?

Clinical heterogeneity

Sometimes trials are just looking at different concepts. Reviewers might set out to summarise interventions for improving patients' ability to make treatment choices; the trials, however, might have covered diverse interventions, such as information leaflets, CD Roms, counselling sessions with a nurse, and training in consultation techniques for doctors. Although the interventions try to achieve the same end result (to improve patients' ability to make choices), they are different in nature.

In theory, we could add all the trials in this review together and come up with a number, but would this be useful? Would the averaged number apply to all these diverse interventions? The interventions are so different that combining them does not make clinical sense. This is an example of clinical heterogeneity. Other circumstances that may give rise to clinical heterogeneity include differences in selection of patients, severity of disease, and management. Judgments about clinical heterogeneity are qualitative, do not involve any calculations, and can be made by putting forward a convincing argument about similarities (or differences) between the trials.

Statistical heterogeneity

Individual trials in a systematic review may seem to measure the same outcome but may have results that are not consistent with each other. Some trials show a benefit while others show harm, or the trials are inconsistent in the size of benefit or harm. This is the case in the systematic review of medications to prevent allergic reactions caused by contrast media.¹ The trials that measured effects on cutaneous symptoms of allergy showed a range of odds ratios from 0.12 favouring the medication to 1.02 favouring the control (fig 1). This is an example of statistical heterogeneity.

How can you detect it and does it matter?

Statistical heterogeneity is apparent only after the analysis of the results. Heterogeneity may be judged graphically (by looking at the forest plot) and be measured statistically. In a forest plot from the systematic review of calcium supplementation,² the error bars for each trial include the summary result, which suggests that statistical heterogeneity is not a problem and that the message is a consistent one (fig 2).

To determine whether significant heterogeneity exists, look for the P value for the χ^2 test of heterogeneity. A high P value is good news because it suggests that the heterogeneity is insignificant and that one can go ahead and summarise the results. Because statistical tests for heterogeneity are not very powerful it is sensible to use a higher P value than usual (say, $P > 0.1$) as the cut-off for a decision and to think about clinical heterogeneity anyway.

The systematic review of calcium supplementation passes the test, and the authors have rightly summarised the effects on bone density using a simple fixed effects model. This model assumes that all trials

RESEARCH p82

clinical epidemiologist
BMJ, London WC1H 9JR
jfletcher@bmj.com

BMJ 2007;334:94-6

doi: 10.1136/bmj.39057.406644.68

This is the first in a series of occasional articles explaining statistical and epidemiological tests used in research papers in the *BMJ*.

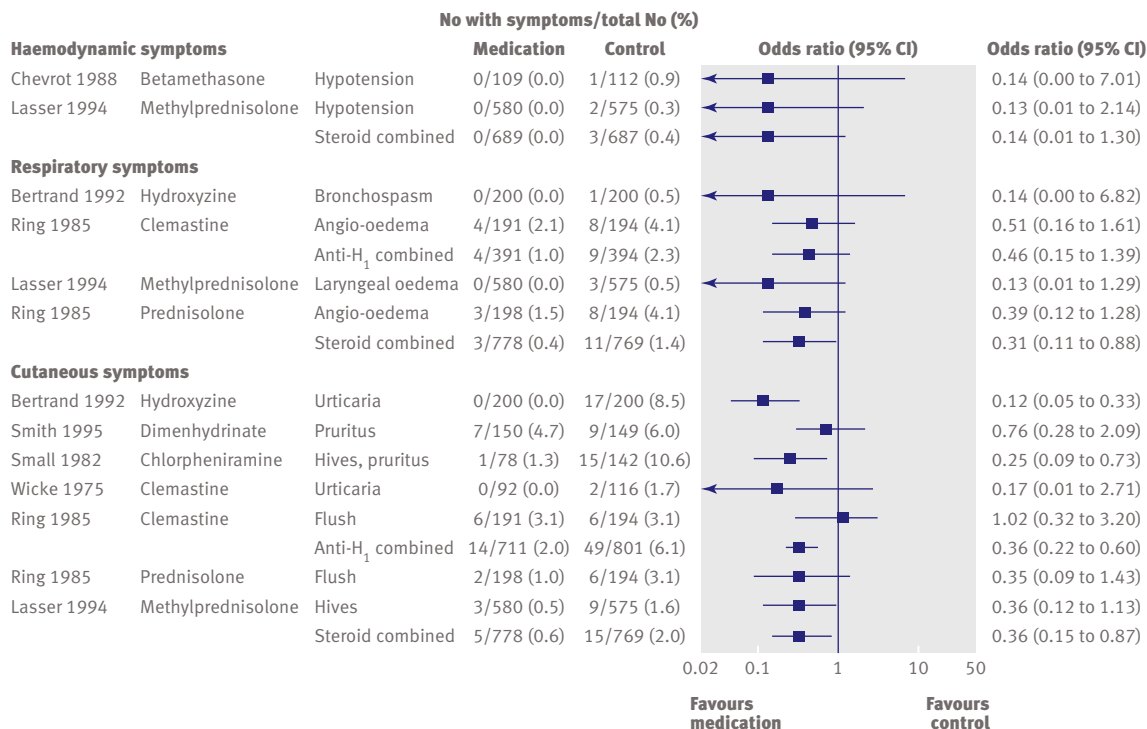


Fig 1 | Forest plot adapted from Tramèr et al¹ showing statistical heterogeneity in the odds ratios for medications to prevent cutaneous allergic reactions (P for 2 test for heterogeneity for anti-H1 combined was 0.03)

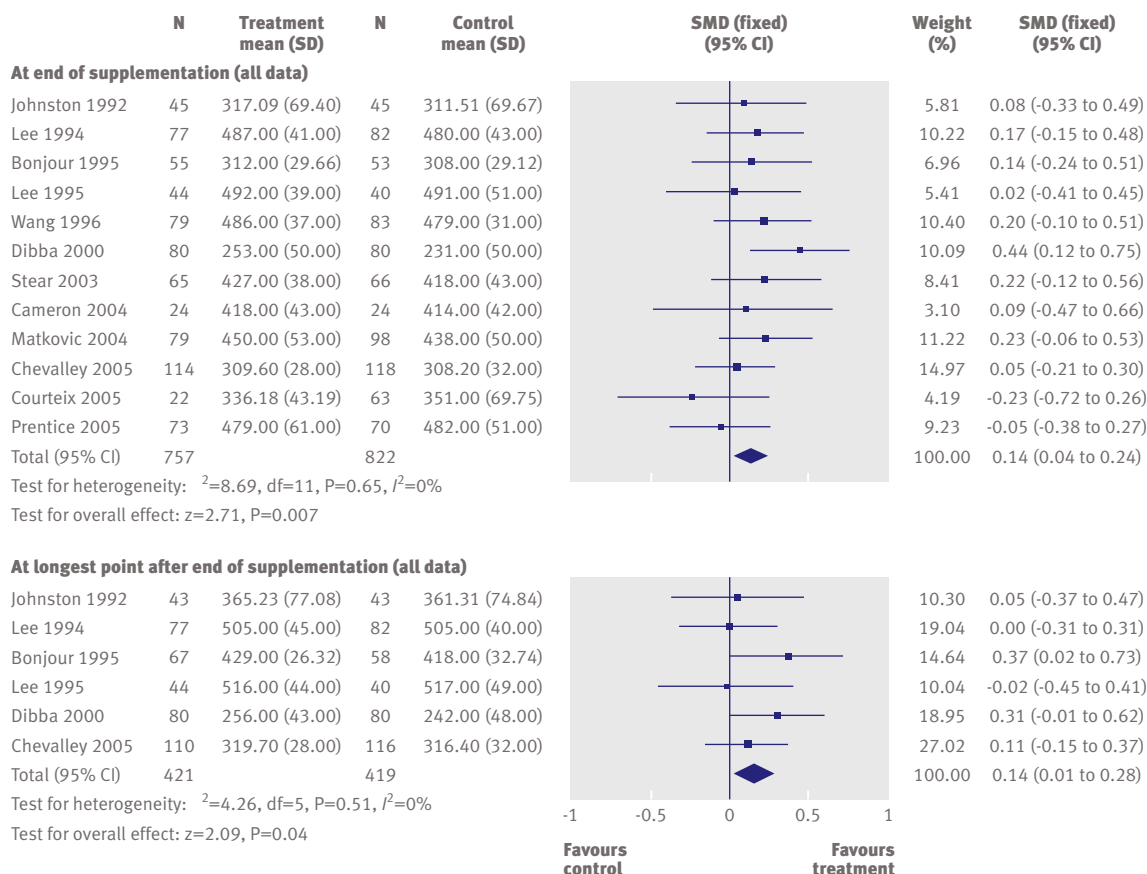


Fig 2 | Forest plot adapted from Winzenberg et al² showing absence of statistical heterogeneity in the odds ratios for the effect of calcium supplementation on bone mineral density. SMD=standardised mean difference

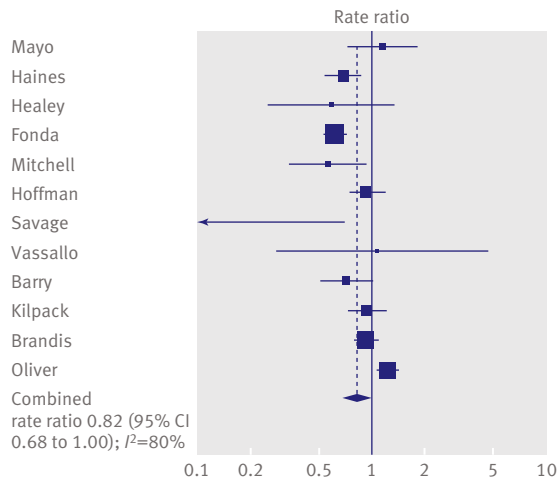


Fig 3 | Forest plot from Oliver et al³ showing rate ratios (random effects model) for the effects of strategies to prevent falls

are trying to measure the same thing and that more influence should be given to larger trials when computing an average effect.⁴

But what if the P value for the χ^2 test of heterogeneity is low, suggesting significant heterogeneity? What can be done? Two approaches are possible. We can either avoid summarising the result and look for reasons for the heterogeneity, or we can summarise the effects using another method—the random effects model. Reasons for heterogeneity, other than clinical differences, could include methodological issues such as problems with randomisation, early termination of trials, use of absolute rather than relative measures of risk, and publication bias.

The authors of the systematic review of medications used to prevent allergic reactions caused by contrast media took the first approach.¹ The forest plots suggest that the two classes of drugs have different effects, particularly for skin reactions, and the P value for the statistical test for heterogeneity was significant at 0.03. They decided not to summarise an average effect and felt that the difference between treatments was part of the message of the review.

The authors of the review of interventions to prevent falls and fractures took the second approach.³ The forest plot for falls in hospital shows a wide spread of results (fig 3). Some trials suggest benefit and others suggest harm from the multifaceted interventions. The authors present the I^2 statistic, which measures the percentage of variation that is not due to chance. A high percentage, such as the 80% seen here, suggests important heterogeneity. (An I^2 value of <25% is considered low.)⁵

Nevertheless, the authors felt that all the trials were trying to measure essentially the same thing and that it was worth summarising the results. They used the random effects model, which uses a different formula to calculate more conservative 95% confidence intervals. The effects of treatment are assumed to vary around some overall average treatment effect, as opposed to a fixed effects model, in which it is assumed that each study has the same fixed common treatment effect.⁴

FURTHER READING

- Chalmers I, Altman DG. *Systematic reviews*. London: BMJ Publishing, 1995.
- Thompson SG. Why sources of heterogeneity in meta-analyses should be investigated. *BMJ* 1994;309:1351-5.

USEFUL QUESTIONS TO CONSIDER

- Was it really a good idea to combine the trials?
- Is there too much clinical heterogeneity for the review to make sense?
- Do the forest plots look consistent?
- Do the statistical tests suggest that heterogeneity is a problem?

Systematic reviews with a meta-analysis try to provide better numerical answers to the questions, “what is the effect of this intervention and how sure are we about that?” But before believing the results of this method, it might be useful to consider four questions (see box).

Contributors: JF is the sole contributor.

Competing interests: None declared.

1. Tramèr M, von Elm E, Loubeyre P, Hauser C. Pharmacological prevention of serious anaphylactic reactions due to iodinated contrast media: systematic review. *BMJ* 2006;333:675-8.
2. Winzenberg T, Shaw K, Fryer J, Jones G. Effects of calcium supplementation on bone density in healthy children: meta-analysis of randomised controlled trials. *BMJ* 2006;333:775-8.
3. Oliver D, Connelly JB, Victor CR, Shaw FE, Whitehead A, Genc Y, et al. Strategies to prevent falls and fractures in hospitals and care homes and effect of cognitive impairment: systematic review and meta-analyses. *BMJ* 2007 doi: 10.1136/bmj.39049.706493.55
4. Higgins JPT, Green S. Summarising effects across studies. *Cochrane handbook for systematic reviews of interventions* 4.2.6 [updated Sep 2006]; Section 8.6. In: The Cochrane Library, Issue 4. Chichester: Wiley, 2006. www.cochrane.org/resources/handbook
5. Higgins J, Thompson S, Deeks J, Altman D. Measuring inconsistency in meta-analyses *BMJ* 2003;327:557-60.

CORRECTIONS AND CLARIFICATIONS

Etoricoxib and diclofenac are associated with similar cardiovascular risks

A momentary lapse in concentration led to a transposition of data when reporting the findings of a recent study in this Short Cuts item by Alison Tonks (*BMJ* 2006;333:1113, 25 Nov, doi: 10.1136/bmj.333.7578.1113-a). The event rates (of thrombotic cardiovascular events such as heart attack) according to the pooled analysis from three large clinical trials in patients with arthritis should have been given as 1.30 [not 1.24] per 100 patient years for diclofenac and 1.24 [not 1.30] for etoricoxib. The hazard ratio was correct: 0.95 (95% confidence interval 0.81 to 1.11) for etoricoxib vs diclofenac, with 0.81 being etoricoxib.

GMC strikes off expert in drug addiction

In this news article by Owen Dyer we wrongly stated that Colin Brewer “took over the Stapleford Centre [. . .] in 1987 after the clinic’s previous director was found guilty of overprescribing” (*BMJ* 2006;333:1035, 18 Nov, doi: 10.1136/bmj.39034.335278.DB). In fact, it was Brewer who set up the Stapleford Centre, so the centre had no previous director.

Preventing and treating hepatitis B infection

Two years on, we have been alerted to an error in a box in this clinical review by Rakesh Aggarwal and Piyush Ranjan (*BMJ* 2004;329:1080-6, doi: 10.1136/bmj.329.7474.1080). The error occurs only in the “Full Text” (html) version of the article (not in the pdf or the printed journal). In box 5, a vital superscript value was missing: the second bullet point should read “Virological response—Decline in hepatitis B virus DNA to <10⁵ copies/ml.” We have already published these corrections on bmj.com