Measuring inconsistency in meta-analyses

Julian P T Higgins, Simon G Thompson, Jonathan J Deeks, Douglas G Altman

Cochrane Reviews have recently started including the quantity I^2 to help readers assess the consistency of the results of studies in meta-analyses. What does this new quantity mean, and why is assessment of heterogeneity so important to clinical practice?

MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 2SR
Julian P T Higgins *statistician*
Simon G Thompson *director*
Cancer Research UK/NHS Centre for Statistics in Medicine, Institute of Health Sciences, Oxford OX3 7LF
Jonathan J Deeks *senior medical statistician*
Douglas G Altman *professor of statistics in medicine*
Correspondence to: J P T Higgins julian.higgins@mrc-bsu.cam.ac.uk

BMJ 2003;327:557–60

Systematic reviews and meta-analyses can provide convincing and reliable evidence relevant to many aspects of medicine and health care.[1] Their value is especially clear when the results of the studies they include show clinically important effects of similar magnitude. However, the conclusions are less clear when the included studies have differing results. In an attempt to establish whether studies are consistent, reports of meta-analyses commonly present a statistical test of heterogeneity. The test seeks to determine whether there are genuine differences underlying the results of the studies (heterogeneity), or whether the variation in findings is compatible with chance alone (homogeneity). However, the test is susceptible to the number of trials included in the meta-analysis. We have developed a new quantity, I^2, which we believe gives a better measure of the consistency between trials in a meta-analysis.

Need for consistency

Assessment of the consistency of effects across studies is an essential part of meta-analysis. Unless we know how consistent the results of studies are, we cannot determine the generalisability of the findings of the meta-analysis. Indeed, several hierarchical systems for grading evidence state that the results of studies must be consistent or homogeneous to obtain the highest grading.[2–4]

Tests for heterogeneity are commonly used to decide on methods for combining studies and for concluding consistency or inconsistency of findings.[5 6] But what does the test achieve in practice, and how should the resulting P values be interpreted?

Testing for heterogeneity

A test for heterogeneity examines the null hypothesis that all studies are evaluating the same effect. The usual test statistic (Cochran's Q) is computed by summing the squared deviations of each study's estimate from the overall meta-analytic estimate, weighting each study's contribution in the same manner as in the meta-analysis.[7] P values are obtained by comparing the statistic with a χ^2 distribution with k–1 degrees of freedom (where k is the number of studies).

The test is known to be poor at detecting true heterogeneity among studies as significant. Meta-analyses often include small numbers of studies,[6 8] and the power of the test in such circumstances is low.[9 10] For example, consider the meta-analysis of randomised controlled trials of amantadine for preventing influenza (fig 1).[11] The treatment effects in the eight trials seem inconsistent: the reduction in odds vary from 16% to 93%, with some of the confidence intervals not overlapping. But the test of heterogeneity yields a P value of 0.09, conventionally interpreted as being non-significant. Because the test is poor at detecting true heterogeneity, a non-significant result cannot be taken as evidence of homogeneity. Using a cut-off of 10% for significance[12] ameliorates this problem but increases the risk of drawing a false positive conclusion (type I error).[10]

Conversely, the test arguably has excessive power when there are many studies, especially when those studies are large. One of the largest meta-analyses in the *Cochrane Database of Systematic Reviews* is of clinical trials of tricyclic antidepressants and selective serotonin reuptake inhibitors for treatment of depression.[13] Over 15 000 participants from 135 trials are included in the assessment of comparative drop-out rates, and the test for heterogeneity is significant (P = 0.005). However, this P value does not reasonably describe the extent of heterogeneity in the results of the trials. As we show later, a little inconsistency exists among these trials but it does not affect the conclusion of the review (that serotonin reuptake inhibitors have lower discontinuation rates than tricyclic antidepressants).

Since systematic reviews bring together studies that are diverse both clinically and methodologically, heterogeneity in their results is to be expected.[6] For example, heterogeneity is likely to arise through diversity in doses, lengths of follow up, study quality, and inclusion criteria for participants. So there seems little point in simply testing for heterogeneity when what matters is the extent to which it affects the conclusions of the meta-analysis.

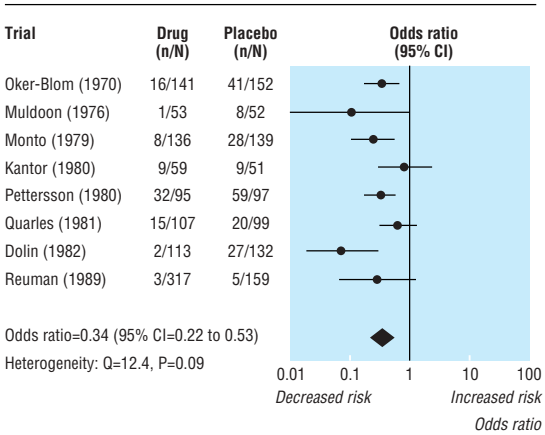

Fig 1 Eight trials of amantadine for prevention of influenza.[11] Outcome is cases of influenza. Summary odds ratios calculated with random effects method

Table 1 Heterogeneity statistics for examples of meta-analyses from the literature. Meta-analyses were conducted using either *meta* or *metan* in STATA[15]

Topic	Outcome/analysis	Effect measure	No of studies	Heterogeneity test			I^2 (95% uncertainty interval)*
				Q	df	P	
Tamoxifen for breast cancer[16]	Mortality	Peto odds ratio	55	55.9	54	0.40	3 (0 to 28)
Streptokinase after myocardial infarction[17]	Mortality	Odds ratio	33	39.5	32	0.17	19 (0 to 48)
Selective serotonin reuptake inhibitors for depression[13]	Drop-out	Odds ratio	135	179.9	134	0.005	26 (7 to 40)
Magnesium for acute myocardial infarction[18]	Death	Odds ratio	16	40.2	15	0.0004	63 (30 to 78)
Magnetic fields and leukaemia[19]	All studies	Odds ratio	6	15.9	5	0.007	69 (26 to 87)
Amantadine[11]	Prevention of influenza	Odds ratio	8	12.44	7	0.09	44 (0 to 75)

df=degrees of freedom.
*Values of I^2 are percentages. 95% uncertainty intervals are calculated as proposed by Higgins and Thompson.[14]

Quantifying heterogeneity: a better approach

We developed an alternative approach that quantifies the effect of heterogeneity, providing a measure of the degree of inconsistency in the studies' results.[14] The quantity, which we call I^2, describes the percentage of total variation across studies that is due to heterogeneity rather than chance. I^2 can be readily calculated from basic results obtained from a typical meta-analysis as $I^2 = 100\% \times (Q - \mathrm{df})/Q$, where Q is Cochran's heterogeneity statistic and df the degrees of freedom. Negative values of I^2 are put equal to zero so that I^2 lies between 0% and 100%. A value of 0% indicates no observed heterogeneity, and larger values show increasing heterogeneity.

Examples of values of I^2

The principal advantage of I^2 is that it can be calculated and compared across meta-analyses of different sizes, of different types of study, and using different types of outcome data. Table 1 gives I^2 values for six published meta-analyses along with 95% uncertainty intervals. The upper limits of these intervals show that conclusions of homogeneity in meta-analyses of small numbers of studies are often unjustified.[11 13 15–19]

The tamoxifen and streptokinase meta-analyses, in which all the included studies found similar effects,[16 17] have I^2 values of 3% and 19% respectively. These indicate little variability between studies that cannot be explained by chance. For the review comparing drop-outs on selective serotonin reuptake inhibitors with tricyclic antidepressants, I^2 is 26%, indicating that although the heterogeneity is highly significant, it is a small effect.

The reviews of trials of magnesium after myocardial infarction ($I^2 = 63\%$) and case-control studies investigating the effects of electromagnetic radiation on leukaemia (69%) both included studies with diverse results. The high I^2 values show that most of the variability across studies is due to heterogeneity rather than chance. Although no significant heterogeneity was detected in the review of amantadine,[11] the inconsistency was moderately large ($I^2 = 44\%$).

Figure 2 shows the observed values of I^2 from 509 meta-analyses in the *Cochrane Database of Systematic Reviews*. Almost half of these meta-analyses (250) had no inconsistency ($I^2 = 0\%$). Among meta-analyses with some heterogeneity, the distribution of I^2 is roughly flat.

Further applications of I^2

I^2 can also be helpful in investigating the causes and type of heterogeneity, as in the three examples below.

Methodological subgroups

Figure 3 shows the six case-control studies of magnetic fields and leukaemia broken down into two subgroups based on assessment of their quality.[19] If heterogeneity is identified in a meta-analysis a common option is to subgroup the studies. Because of loss of power, non-significant heterogeneity within a subgroup may be due not to homogeneity but to the smaller number of studies. Here, the P values for the heterogeneity test are higher for the two subgroups (P = 0.3 and P = 0.009) than for the complete data (P = 0.007), which suggests greater consistency within the subgroups. However, the values of I^2 show that the three low quality studies are more inconsistent ($I^2 = 79\%$) than all six ($I^2 = 69\%$) (table 2). Substantially less inconsistency exists among the high quality studies ($I^2 = 15\%$), although uncertainty intervals for all of the I^2 values are wide.

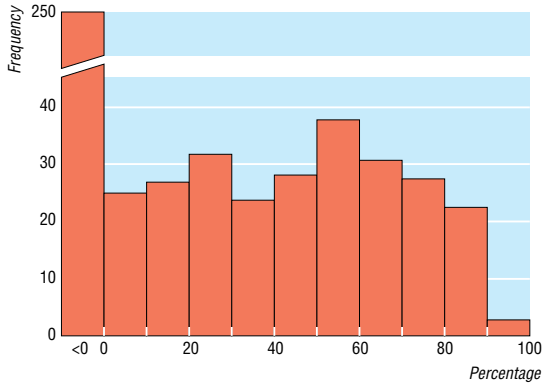

Fig 2 Distribution of observed values of I^2 based on odds ratios from 509 meta-analyses of dichotomous outcomes in the *Cochrane Database of Systematic Reviews*. Data are from the first subgroup (if any) in the first meta-analysis (if any) in each review, if it involved a dichotomous outcome and at least two trials with events. Meta-analyses conducted with *metan* in STATA[15]

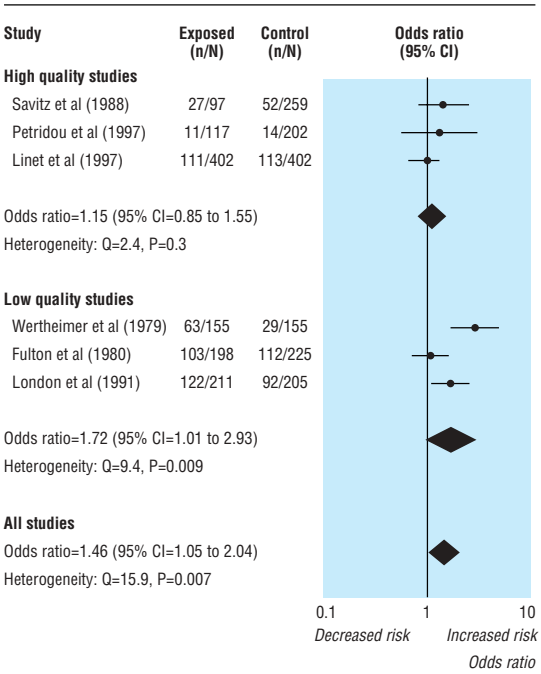

Fig 3 Meta-analyses of six case-control studies relating residential exposure to electromagnetic fields to childhood leukaemia.[19] Summary odds ratio calculated by random effects method

Heterogeneity related to choice of effect measure

A systematic review of clinical trials of human albumin administration in critically ill patients concluded that albumin may increase mortality.[20] These studies had no inconsistency in risk ratio estimates ($I^2=0\%$) and a narrow uncertainty interval. Table 2 shows the heterogeneity statistics for risk differences as well as for risk ratios. Six trials with no deaths in either treatment group do not contribute information on risk ratios, but they all provide estimates of risk differences. Using P values to decide which scale is more consistent with the data[21] is inappropriate because of the differing numbers of studies. I^2 values may validly be compared and show that the risk differences are less homogeneous, as is often the case.[22]

Clinically important subgroups

I^2 can also be used to describe heterogeneity among subgroups. Table 2 includes results for the outcome of recurrence in the meta-analysis of trials of tamoxifen for women with early breast cancer. There was highly significant (P = 0.00002) and important heterogeneity ($I^2=50\%$) among the trials.[16] However, a potentially important source of heterogeneity is the duration of treatment. The authors divided the trials into three duration categories and presented an overall heterogeneity test, a test comparing the three subgroups, and a test for heterogeneity within the subgroups. I^2 values corresponding to each test show that 96% of the variability observed among the three subgroups cannot be explained by chance. This is not clear from the P values alone. The extreme inconsistency among all 55 trials in the odds ratios for recurrence ($I^2=50\%$) is substantially reduced ($I^2=13\%$) once differences in treatment duration are accounted for.

How much is too much heterogeneity?

A naive categorisation of values for I^2 would not be appropriate for all circumstances, although we would tentatively assign adjectives of low, moderate, and high to I^2 values of 25%, 50%, and 75%. Figure 2 shows that about a quarter of meta-analyses have I^2 values over 50%. Quantification of heterogeneity is only one component of a wider investigation of variability across studies, the most important being diversity in clinical and methodological aspects. Meta-analysts must also consider the clinical implications of the observed degree of inconsistency across studies. For example, interpretation of a given degree of heterogeneity across several studies will differ according to whether the estimates show the same direction of effect.

Advantages of I^2

- Focuses attention on the effect of any heterogeneity on the meta-analysis
- Interpretation is intuitive—the percentage of total variation across studies due to heterogeneity
- Can be accompanied by an uncertainty interval
- Simple to calculate and can usually be derived from published meta-analyses
- Does not inherently depend on the number of studies in the meta-analysis
- May be interpreted similarly irrespective of the type of outcome data (eg dichotomous, quantitative, or time to event) and choice of effect measure (eg odds ratio or hazard ratio)
- Wide range of applications

Table 2 More advanced applications of I^2 for assessing heterogeneity in three published meta-analyses. Meta-analyses were conducted with either *meta* or *metan* in STATA[15]

Topic	Outcome/analysis	Effect measure	No of studies	Heterogeneity test			I^2 (95% uncertainty intervals)*
				Q	df	P	
Magnetic fields and leukaemia[19]	All studies	Odds ratio	6	15.9	5	0.007	69 (26 to 87)
	High quality	Odds ratio	3	2.4	2	0.31	17 (0 to 91)
	Low quality	Odds ratio	3	9.4	2	0.009	79 (32 to 94)
Human albumin for critically ill[20]	Death	Risk ratio	24†	15.3	23	0.88	0 (0 to 17)
	Death	Risk difference	30	36.7	29	0.15	21 (0 to 50)
Tamoxifen to prevent recurrence of breast cancer[17]	All studies	Peto odds ratio	55	108.2	54	0.00002	50 (32 to 63)
	Total within groups‡	Peto odds ratio	—	59.9	52	0.21	13 (0 to 39)
	Between groups‡	Peto odds ratio	3 groups	48.3	2	<0.00001	96 (91 to 98)

df=degrees of freedom,
*Values of I^2 are percentages. 95% uncertainty intervals are calculated as proposed by Higgins and Thompson.[14]
†Studies with no events in either treatment group do not contribute to this analysis.
‡Subgroup defined by duration of tamoxifen treatment.

Summary points

Inconsistency of studies' results in a meta-analysis reduces the confidence of recommendations about treatment

Inconsistency is usually assessed with a test for heterogeneity, but problems of power can give misleading results

A new quantity I^2, ranging from 0-100%, is described that measures the degree of inconsistency across studies in a meta-analysis

I^2 can be directly compared between meta-analyses with different numbers of studies and different types of outcome data

I^2 is preferable to a test for heterogeneity in judging consistency of evidence

An alternative quantification of heterogeneity in a meta-analysis is the among-study variance (often called τ^2), calculated as part of a random effects meta-analysis. This is more useful for comparisons of heterogeneity among subgroups, but values depend on the treatment effect scale. We believe, I^2 offers advantages over existing approaches to the assessment of heterogeneity (box). Focusing on the effect of heterogeneity also avoids the temptation to perform so called two stage analyses, in which the meta-analysis strategy (fixed or random effects method) is determined by the result of a statistical test. Such strategies have been found to be problematic.[23 24] We therefore believe that I^2 is preferable to the test of heterogeneity when assessing inconsistency across studies.

We thank Keith O'Rourke and Ian White for useful comments.

Contributors: The authors all work as statisticians and have extensive experience in methodological, empirical and applied research in meta-analysis. JH, JD, and DA are coconvenors of the Cochrane Statistical Methods Group. The views expressed in the paper are those of the authors. All authors contributed to the development of the methods described. JH and ST worked more closely on the development of I^2. JH is guarantor.

Funding: This work was funded in part by MRC Project Grant G9815466.

Competing interests: None declared.

1 Egger M, Davey Smith G. Meta-analysis: potentials and promise. *BMJ* 1997;315:1371-4.
2 Liberati A, Buzzetti R, Grilli R, Magrini N, Minozzi S. Which guidelines can we trust? *West J Med* 2001;174:262-5.
3 Harbour R, Miller J for the Scottish Intercollegiate Guidelines Network Grading Review Group. A new system for grading recommendations in evidence based guidelines. *BMJ* 2001;323:334-6.
4 Guyatt G, Sinclair J, Cook D, Jaeschke R, Schünemann H, Pauker S. Moving from evidence to action. In: Guyatt G, Rennie D, eds. *Users' guides to the medical literature: a manual for evidence-based clinical practice*. Chicago: American Medical Association, 2002:599-608.
5 Petitti DB. Approaches to heterogeneity in meta-analysis. *Stat Med* 2001;20:3625-33.
6 Higgins J, Thompson S, Deeks J, Altman D. Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. *J Health Serv Res Policy* 2002;7:51-61.
7 Cochran WG. The combination of estimates from different experiments. *Biometrics* 1954;10:101-29.
8 Sterne JAC, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol* 2001;54:1046-55.
9 Paul SR, Donner A. Small sample performance of tests of homogeneity of odds ratios in k 2×2 tables. *Stat Med* 1992;11:159-65.
10 Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Stat Med* 1998;17:841-56.
11 Jefferson TO, Demicheli V, Deeks JJ, Rivetti D. Amantadine and rimantadine for preventing and treating influenza A in adults. *Cochrane Database Syst Rev* 2002;(4):CD001169.
12 Dickersin K, Berlin JA. Meta-analysis: state-of-the-science. *Epidemiol Rev* 1992;14:154-76.
13 Barbui C, Hotopf M, Freemantle N, Boynton J, Churchill R, Eccles MP, Geddes JR, et al. Treatment discontinuation with selective serotonin reuptake inhibitors (SSRIs) versus tricyclic antidepressants (TCAs). *Cochrane Database Syst Rev* 2003;(3):CD002791.
14 Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21:1539-58.
15 Sterne JAC, Bradburn MJ, Egger M. Meta-analysis in STATA. In: Egger M, Davey Smith G, Altman DG, eds. *Systematic reviews in health care: meta-analysis in context*. 2nd ed. London: BMJ Publications, 2001:347-69.
16 Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet* 1998;351:1451-67.
17 Lau J, Antman EM, Jimenez-Silva J, Kupelink B, Mosteller SF, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N Engl J Med* 1992;327:248-54.
18 Egger M, Davey Smith G. Misleading meta-analysis. *BMJ* 1995;310:752-4.
19 Angelillo IF, Villari P. Residential exposure to electromagnetic fields and childhood leukaemia: a meta-analysis. *Bull World Health Organ* 1999;77:906-15.
20 Cochrane Injuries Group Albumin Reviewers. Human albumin administration in critically ill patients: systematic review of randomised controlled trials. *BMJ* 1998;317:235-40.
21 Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Stat Med* 2000;19:1707-28.
22 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.
23 Freeman PR. The performance of the two-stage analysis of two-treatment, two-period crossover trials. *Stat Med* 1989;8:1421-32.
24 Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J Clin Epidemiol* 1999;52:935-42.

(Accepted 16 June 2003)

One hundred years ago

The open-air treatment of surgical tuberculosis

The profound and far-reaching effects of what is popularly known as the open-air treatment of consumption are only just beginning to be appreciated. Having originated merely with an attempt to cure or alleviate a disease which afflicts every race of mankind, it has directed attention to the physical and social conditions which lead to it, and clearly indicates the direction in which reform is urgently needed if deterioration of the physique of modern urban communities is to be arrested.

What, then, is this great discovery embodied in the open-air treatment of consumption? Absurd as it may sound, it is nothing but the rediscovery of the *vis medicatrix Naturae* and of the value of unpolluted air. That the body possesses a certain power of recovering from illness and repairing wounds, and that pure air is beneficial to health, have always been familiar facts; but familiar facts are just those which are most constantly disregarded in practice. It is one of the great merits of the open-air treatment of consumption that it is rapidly popularizing the conception that polluted air is as much to be avoided as polluted water. More important still, has been the effect of the open-air treatment in refuting deeply-rooted superstitions as to the evil effects of exposure to atmospheric changes, and in directing attention to the real enemy—namely, dust and dirt, especially the organic dirt which emanates from the animal body.

There can be no great progress in public health until all classes recognize that the first essential of health is minute cleanliness of body, raiment, food, and dwelling-house. How far we are from this ideal, even among the well-to-do classes, every doctor knows.

(*BMJ* 1903;ii:986)