Systematic reviews and meta-analyses can provide convincing and reliable evidence relevant to many aspects of medicine and health care. 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, I², 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. Tests for heterogeneity are commonly used to decide on methods for combining studies and for concluding consistency or inconsistency of findings. 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. P values are obtained by comparing the statistic with a χ² 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, and the power of the test in such circumstances is low. For example, consider the meta-analysis of randomised controlled trials of amantadine for preventing influenza (fig 1). 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 ameliorates this problem but increases the risk of drawing a false positive conclusion (type I error).

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. 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. 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.
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. 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 \( I^2 = 100\%×(Q − df)/Q \) 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. The tamoxifen and streptokinase meta-analyses, in which all the included studies found similar effects, 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, 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. 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.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Outcome/analysis</th>
<th>Effect measure</th>
<th>No of studies</th>
<th>Heterogeneity test</th>
<th>( I^2 ) (95% uncertainty interval)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tamoxifen for breast cancer</td>
<td>Mortality</td>
<td>Peto odds ratio</td>
<td>55</td>
<td>Q=55.9 df=54 P=0.40</td>
<td>3 (0 to 28)</td>
</tr>
<tr>
<td>Streptokinase after myocardial infarction</td>
<td>Mortality</td>
<td>Odds ratio</td>
<td>33</td>
<td>Q=39.5 df=32 P=0.17</td>
<td>19 (0 to 48)</td>
</tr>
<tr>
<td>Selective serotonin reuptake inhibitors for depression</td>
<td>Drop-out</td>
<td>Odds ratio</td>
<td>135</td>
<td>Q=179.9 df=134 P=0.006</td>
<td>26 (7 to 40)</td>
</tr>
<tr>
<td>Magnesium for acute myocardial infarction</td>
<td>Death</td>
<td>Odds ratio</td>
<td>16</td>
<td>Q=40.2 df=15 P=0.0004</td>
<td>63 (30 to 78)</td>
</tr>
<tr>
<td>Magnetic fields and leukemia</td>
<td>All studies</td>
<td>Odds ratio</td>
<td>8</td>
<td>Q=15.9 df=5 P=0.007</td>
<td>69 (26 to 87)</td>
</tr>
<tr>
<td>Amantadine</td>
<td>Prevention of influenza</td>
<td>Odds ratio</td>
<td>8</td>
<td>Q=12.4 df=7 P=0.09</td>
<td>44 (0 to 75)</td>
</tr>
</tbody>
</table>

\( * \) = degrees of freedom.

Values of \( I^2 \) are percentages. 95% uncertainty intervals are calculated as proposed by Higgins and Thompson.14
inconsistency in risk ratio estimates (I²).

Values of

df = degrees of freedom, values to decide which scale is more consistent with the group do not contribute information on risk ratios, but geneity statistics for risk differences as well as for risk

narrow uncertainty interval. Table 2 shows the hetero-

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 (I² = 50%) is substantially reduced (I² = 13%) once differences in treatment duration are accounted for.

How much is too much heterogeneity?

A naive categorisation of values for I² would not be appropriate for all circumstances, although we would tentatively assign adjectives of low, moderate, and high to I² values of 25%, 50%, and 75%. Figure 2 shows that about a quarter of meta-analyses have I² 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.

Table 2

<table>
<thead>
<tr>
<th>Topic</th>
<th>Outcome/analysis</th>
<th>Effect measure</th>
<th>No of studies</th>
<th>Heterogeneity test</th>
<th>I² (95% uncertainty intervals)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q</td>
<td>df</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>Magnetic fields and leukemia†</td>
<td>All studies</td>
<td>Odds ratio</td>
<td>6</td>
<td>15.9</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>High quality</td>
<td>Odds ratio</td>
<td>3</td>
<td>2.4</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Low quality</td>
<td>Odds ratio</td>
<td>3</td>
<td>9.4</td>
<td>2</td>
</tr>
<tr>
<td>Human albumin for critically ill‡</td>
<td>Death</td>
<td>Risk ratio</td>
<td>241</td>
<td>15.3</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Death</td>
<td>Risk difference</td>
<td>30</td>
<td>36.7</td>
<td>29</td>
</tr>
<tr>
<td>Tamoxyfen to prevent recurrence of breast cancer‡</td>
<td>All studies</td>
<td>Peto odds ratio</td>
<td>55</td>
<td>108.2</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>Total within groups</td>
<td>Peto odds ratio</td>
<td>55</td>
<td>59.9</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>Between groups‡</td>
<td>Peto odds ratio</td>
<td>3 groups</td>
<td>46.3</td>
<td>2</td>
</tr>
</tbody>
</table>

df = degrees of freedom.

Values of I² 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.

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 (I² = 0%) and a narrow uncertainty interval. Table 2 shows the hetero-

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 (I² = 0%) and a narrow uncertainty interval. Table 2 shows the hetero-

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 (I² = 0%) and a narrow uncertainty interval. Table 2 shows the hetero-

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 (I² = 0%) and a narrow uncertainty interval. Table 2 shows the hetero-

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 (I² = 0%) and a narrow uncertainty interval. Table 2 shows the hetero-

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 (I² = 0%) and a narrow uncertainty interval. Table 2 shows the hetero-

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.
The open-air treatment of surgical tuberculosis

One hundred years ago

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.

Funding: This work was funded in part by MRC Project Grant G9815-666.

Competing interests: None declared.


400 years ago

The open-air treatment of surgical tuberculosis

One hundred years ago

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.

Funding: This work was funded in part by MRC Project Grant G9815-666.

Competing interests: None declared.


(Accepted 16 June 2003)