

# Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses<sup>†</sup>

Eric A. Engels<sup>1,\*</sup>, Christopher H. Schmid<sup>1,2</sup>, Norma Terrin<sup>1,2</sup>, Ingram Olkin<sup>3</sup>,  
and Joseph Lau<sup>1</sup>

<sup>1</sup> *Division of Clinical Care Research, Department of Medicine, New England Medical Center,  
Tufts University School of Medicine, Boston, MA 02111, U.S.A.*

<sup>2</sup> *Biostatistics Research Center, New England Medical Center, Tufts University School of Medicine, Boston, MA 02111, U.S.A.*

<sup>3</sup> *Department of Statistics, Sequoia Hall, Stanford University, Stanford, CA 94305, U.S.A.*

## SUMMARY

For meta-analysis, substantial uncertainty remains about the most appropriate statistical methods for combining the results of separate trials. An important issue for meta-analysis is how to incorporate heterogeneity, defined as variation among the results of individual trials beyond that expected from chance, into summary estimates of treatment effect. Another consideration is which 'metric' to use to measure treatment effect; for trials with binary outcomes, there are several possible metrics, including the odds ratio (a relative measure) and risk difference (an absolute measure). To examine empirically how assessment of treatment effect and heterogeneity may differ when different methods are utilized, we studied 125 meta-analyses representative of those performed by clinical investigators. There was no meta-analysis in which the summary risk difference and odds ratio were discrepant to the extent that one indicated significant benefit while the other indicated significant harm. Further, for most meta-analyses, summary odds ratios and risk differences agreed in statistical significance, leading to similar conclusions about whether treatments affected outcome. Heterogeneity was common regardless of whether treatment effects were measured by odds ratios or risk differences. However, risk differences usually displayed more heterogeneity than odds ratios. Random effects estimates, which incorporate heterogeneity, tended to be less precisely estimated than fixed effects estimates. We present two exceptions to these observations, which derive from the weights assigned to individual trial estimates. We discuss the implications of these findings for selection of a metric for meta-analysis and incorporation of heterogeneity into summary estimates. Published in 2000 by John Wiley & Sons, Ltd.

## 1. INTRODUCTION

Increasingly, meta-analysis is used to synthesize results from randomized controlled trials in clinical medicine. The number of published meta-analyses has grown exponentially [1], and their

---

\* Correspondence to: Eric A. Engels, Viral Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd, MSC 7248, Rockville, MD 20852, U.S.A.

<sup>†</sup> This article is a U.S. Government work and is in the public domain in the U.S.A.

Contract/grant sponsor: Agency for Healthcare Research and Quality; contract/grant numbers: 290970019, T32 HS00060, R01 HS08532, HS10064

Contract/grant sponsor: National Center for Research Resources; contract/grant number: R41 RR12416

potential to change patient care is clearly established [2]. Although application of meta-analytic techniques to clinical problems generates enthusiasm, substantial uncertainty remains about the most appropriate methods for combining the results of separate trials.

Heterogeneity, by which we mean variation among the results of individual trials beyond that expected from chance alone, is an important issue in meta-analysis. Heterogeneity may indicate that trials evaluated different interventions or different populations. When heterogeneity is present, it may be inappropriate to combine the separate trial estimates into a single number, particularly using fixed effects methods that assume a common treatment effect. Random effects methods, which provide an attractive approach to summarizing heterogeneous results, model heterogeneity as variation of individual trial treatment effects around a population average effect. The key distinction between these two types of models concerns the belief regarding behaviour of trial effects as trial sample sizes get very large. If one believes that the individual trial effects would converge to a common value for all trials, a fixed effects model is appropriate, whereas if one believes that individual trials would still demonstrate separate effects, then a random effects model is preferable. Despite extensive discussion of appropriate analytic approaches to heterogeneity [3–5], no large study has empirically examined how frequently meta-analyses in the medical field are heterogeneous or how often and how much heterogeneity affects results.

Another important consideration for meta-analysis is which measure from individual trials should be used to summarize treatment effects. For trials with binary outcomes, there are several effect measures, or 'metrics', directly available. Simplest of these are the risk difference, which measures absolute treatment effect, and the odds ratio and risk ratio, which measure relative treatment effect. The risk ratio is more easily understood than the odds ratio; however, mathematical advantages of the odds ratio over the risk ratio include its symmetry with respect to 'successes' and 'failures', and the fact that the odds ratio may assume values unrestricted between zero and infinity. Of the risk difference, odds ratio and risk ratio, it can be shown that only the risk difference possesses an unbiased estimator. (I. Olkin, unpublished proof, December 1998). There are also other metrics with desirable mathematical properties (see for example Snedecor and Cochran [6] or Emerson [7]), though these have less direct clinical appeal.

Given these considerations, there may be no metric that is 'best' for all circumstances. The risk difference may be the most relevant metric for clinicians and public policy experts interested in the absolute impact of an intervention in a given population [8, 9]. In contrast, which measure is best suited statistically for summarizing results across several trials and populations in a meta-analysis remains unaddressed. Conclusions drawn from a meta-analysis may depend on which metric is used, especially when measures of absolute or relative benefit vary widely among the trials under consideration, that is, when heterogeneity is present. A percentage change for one metric will not translate to a like percentage change in another metric. In turn, measures of heterogeneity will depend on the metric chosen.

We studied these issues by examining a sample of 125 meta-analyses representative of those performed by clinical investigators. To measure treatment effect for the trials within each meta-analysis, we used the odds ratio and risk difference metrics, both of which have well-established statistical properties and are used in published meta-analyses. For each meta-analysis, we derived heterogeneity measures and fixed and random effects summary estimates. We addressed empirically three questions that have important implications for meta-analysis. First, how often are the collections of trials used in meta-analysis heterogeneous? Second, when do fixed effects and random effects methods give different estimates of treatment effect? Third, when does summarizing risk differences give a different impression of treatment effect than summarizing

odds ratios? Answers to these questions and a consideration of theoretical issues that they highlight provide a useful framework for evaluating meta-analytic methods.

## 2. METHODS AND MODELS

### 2.1. Search strategy and inclusion criteria

To examine a broadly inclusive set of meta-analyses, we included in the present study meta-analyses obtained systematically from two sources: seven major medical journals that publish relatively large numbers of meta-analyses (1990–1996 issues of *Annals of Internal Medicine*, *Archives of Internal Medicine*, *British Medical Journal*, *Circulation*, *Journal of the American Medical Association*, *Lancet*, and *New England Journal of Medicine*, with meta-analyses identified through a Medline search), and the 1994 Cochrane Pregnancy and Childbirth database (CCPC, a comprehensive database of systematic reviews on perinatal topics) [10].

Included meta-analyses reported randomized controlled trial data as  $2 \times 2$  tables for binary outcomes. To ensure that included meta-analyses had sufficient data to provide valid effect estimates, we required that each have six or more trials with at least one event in the control arm and that the average number of events in the trial control arms be at least five. To avoid very small trials, we excluded from the meta-analysis any trial with fewer than ten subjects in either arm.

### 2.2. Extraction of data

For meta-analyses that examined several outcomes, we included data only for the outcome judged to have the greatest clinical relevance (chosen by J.L., who was blinded to the results of the meta-analysis). Some journal publications included more than one meta-analysis, each conducted on a distinct set of trials; in these cases, we included each meta-analysis separately in the present study.

For any trial in which one arm had zero events,  $1/2$  was added to each cell of the corresponding  $2 \times 2$  table before calculating the statistics described below [11]. Because our inclusion criteria selected meta-analyses that had few trials with arms with zero events, this correction for zero cells had a minimal impact on conclusions.

### 2.3. Meta-analytic summary statistics

**2.3.1. Summary effect estimates for each meta-analysis.** For each meta-analysis we calculated the Mantel–Haenszel odds ratio as the fixed effects summary estimate of the individual trials' common odds ratio [12]. For each trial  $i$ , let  $p_{Ti}$  and  $p_{Ci}$  represent the proportion of subjects with the event in the treatment and control arm, respectively, and  $n_{Ti}$  and  $n_{Ci}$  the corresponding numbers of subjects randomized to each arm. Then the individual trial odds ratio  $OR_i = (p_{Ti})(1 - p_{Ci}) / [(p_{Ci})(1 - p_{Ti})]$ , and the Mantel–Haenszel odds ratio is given by the weighted average  $\sum w_{Oi} OR_i / \sum w_{Oi}$  with weights  $w_{Oi} = [(n_{Ci})(p_{Ci})] [(n_{Ti})(1 - p_{Ti})] / (n_{Ti} + n_{Ci})$ . The standard error of the logarithm of this estimate is discussed by Robins *et al.* [13].

The fixed effects risk difference summary estimate is a weighted average of the trial risk differences  $RD_i = p_{Ti} - p_{Ci}$ , given by  $\sum w_{Ri} RD_i / \sum w_{Ri}$ ; weights are given by  $w_{Ri} =$

$1/[(p_{Ti})(1 - p_{Ti})/n_{Ti} + (p_{Ci})(1 - p_{Ci})/n_{Ci}]$ . These inverse-variance weights minimize the variance of the summary estimate. The variance of the fixed effect risk difference is then  $1/(\sum w_{Ri})$ .

We used the method of DerSimonian and Laird to derive random effects summary estimates for the risk difference [14], and a modification of this method to derive estimates for the odds ratio [15]. The random effects estimates are weighted averages of either the  $RD_i$ 's or the logarithm of the  $OR_i$ 's. The random effects weights  $w_i^*$  are based on the corresponding fixed effects weights  $w_i$ , but also incorporate information, derived from the  $Q$ -statistic described below, regarding heterogeneity among measured trial effects. We thus have

$$w_i^* = 1/(D + 1/w_i)$$

where

$$D = \max\{0, [Q - (K - 1)] (\sum w_i) / [(\sum w_i)^2 - (\sum w_i^2)]\}$$

and where  $K$  is the number of trials. For the random effects risk difference, the fixed effects weights  $w_i$  used in the calculation of  $w_i^*$  are the  $w_{Ri}$ 's. For the logarithm of the odds ratio the fixed effects weights  $w_i$  are given by

$$w_{Li} = 1/\{[n_{Ti}p_{Ti}(1 - p_{Ti})]^{-1} + [n_{Ci}p_{Ci}(1 - p_{Ci})]^{-1}\}$$

The variance of the random effects estimate is  $1/(\sum w_i^*)$ . For the risk differences, it is always true that  $w_i^* \leq w_i$ , so the variance of the fixed effects estimate never exceeds that of the random effects estimate. However, the variance of the Mantel-Haenszel odds ratio estimate can exceed the variance of the random effects odds ratio, because unlike the Mantel-Haenszel odds ratio, the random effects odds ratio is derived as a weighted average of the logarithm of trial odds ratios.

Examination of the weighting formulae suggests how these estimators might differ under various data structures. To simplify matters, we assume that the sample sizes in the two trial treatment arms are approximately equal. This assumption usually holds in the clinical trials we consider in this paper. Then the weights may be rewritten as

$$w_{Ri} = 1/[(p_{Ti})(1 - p_{Ti})/n_{Ti} + (p_{Ci})(1 - p_{Ci})/n_{Ci}] \approx n/[(p_{Ti})(1 - p_{Ti}) + (p_{Ci})(1 - p_{Ci})]$$

$$w_{Oi} = [(n_{Ci})(p_{Ci})] [(n_{Ti})(1 - p_{Ti})]/(n_{Ti} + n_{Ci}) \approx np_{Ci}(1 - p_{Ti})/2$$

$$w_{Li} = 1/\{[n_{Ti}p_{Ti}(1 - p_{Ti})]^{-1} + [n_{Ci}p_{Ci}(1 - p_{Ci})]^{-1}\} \\ \approx n/\{[p_{Ti}(1 - p_{Ti})]^{-1} + [p_{Ci}(1 - p_{Ci})]^{-1}\}$$

when  $n$  is the common size of the trial arms. For comparison here, we also list weights for the Peto estimate of the summary odds ratio [16]:

$$w_{Pi} \approx n\{[p_{Ti} + p_{Ci}][(1 - p_{Ti}) + (1 - p_{Ci})]\}$$

Table I shows the relative weights assigned to individual trial estimates for each summary method, for a hypothetical meta-analysis in which all trials have the same size. Each column of weights is scaled so that the weight assigned to a trial with  $p_T = p_C = 0.50$  would be 1.00. Because of this scaling, weights are easily compared within columns, but direct comparisons of weights between columns are not valid. For example, compared with a trial in which  $p_T = p_C = 0.50$ , the last row of Table I shows that an equally sized trial in which  $p_C = 0.05$  and  $p_T = 0.01$  would receive 8.71 times as much weight in an estimate of the fixed effect summary risk difference

Table I. Relative weights for different estimators.

$p_C$	$p_T$	Risk difference	Odds ratio	$w_R$	$w_O$	$w_L$	$w_P$
0.500	0.500	0.000	1.00	1.00	1.00	1.00	1.00
0.250	0.500	0.250	3.00	1.14	0.50	0.86	0.94
0.250	0.375	0.125	1.80	1.19	0.63	0.83	0.86
0.250	0.250	0.000	1.00	1.33	0.75	0.75	0.75
0.375	0.250	-0.125	0.56	1.19	1.13	0.83	0.86
0.500	0.250	-0.250	0.33	1.14	1.50	0.86	0.94
0.100	0.500	0.400	9.00	1.47	0.20	0.53	0.84
0.100	0.200	0.100	2.25	2.00	0.32	0.46	0.51
0.100	0.150	0.050	1.59	2.30	0.34	0.42	0.44
0.100	0.100	0.000	1.00	2.78	0.36	0.36	0.36
0.150	0.100	-0.050	0.63	2.30	0.54	0.42	0.44
0.200	0.100	-0.100	0.44	2.00	0.72	0.46	0.51
0.500	0.100	-0.400	0.11	1.47	1.80	0.53	0.84
0.050	0.250	0.200	6.33	2.13	0.15	0.30	0.51
0.050	0.100	0.050	2.11	3.64	0.18	0.25	0.28
0.050	0.075	0.025	1.54	4.28	0.19	0.23	0.23
0.050	0.050	0.000	1.00	5.26	0.19	0.19	0.19
0.075	0.050	-0.025	0.65	4.28	0.29	0.23	0.23
0.100	0.050	-0.050	0.47	3.64	0.38	0.25	0.28
0.250	0.050	-0.200	0.16	2.13	0.95	0.30	0.51
0.010	0.050	0.040	5.21	8.71	0.04	0.07	0.12
0.010	0.020	0.010	2.02	16.95	0.04	0.05	0.06
0.010	0.015	0.005	1.51	20.26	0.04	0.05	0.05
0.010	0.010	0.000	1.00	25.25	0.04	0.04	0.04
0.015	0.010	-0.005	0.66	20.26	0.06	0.05	0.05
0.020	0.010	-0.010	0.49	16.95	0.08	0.05	0.06
0.050	0.010	-0.040	0.19	8.71	0.20	0.07	0.12

The rows of the table are ordered in blocks, in descending order with respect to the minimum of  $p_T$  and  $p_C$  within each block.

Abbreviations:  $p_T$ , proportion with outcome in treatment arm;  $p_C$ , proportion with outcome in control arm;  $w_R$ , risk difference weight;  $w_O$ , Mantel-Haenszel odds ratio weight;  $w_L$ , weight for logarithm of odds ratio;  $w_P$ , Peto odds ratio weight.

( $w_R = 8.71$ ), but only one-fifth the weight ( $w_O = 0.20$ ) when using the Mantel-Haenszel method to estimate the odds ratio.

Consideration of the formulae and Table I highlights two findings. First, the risk difference metric gives large weight to trials with small proportions  $p_T$  and  $p_C$ . Second, the various summary methods utilizing the odds ratio metric give large weight to trials with  $p_T$  and  $p_C$  near 0.50.

In calculating the summary effect estimate for an actual meta-analysis, the relative weight given to an individual trial's estimate will also depend on that trial's group sample size  $n$ . Furthermore, for random effects models, the patterns noted in Table I will be mitigated because inclusion of the variance component,  $D$ , tends to make the relative weights more uniform across studies; a trial with large fixed effect weight will receive a relatively smaller weight in a random effects model. The potential effects on the interpretation of meta-analyses that can derive from these differences

in relative weights across metrics and models will be illustrated later by two examples (Sections 3.4.1 and 3.4.2).

*2.3.2. Heterogeneity.* To measure heterogeneity, we calculated  $Q$ -statistics for the trial odds ratios and for the risk differences in each meta-analysis [14].  $Q$  is defined as  $\sum w (Y_i - Y_{\text{fix}})^2$ . For the odds ratios, one uses  $w = w_{Li}$ ,  $Y_i = \log(\text{OR}_i)$ , and  $Y_{\text{fix}} = \sum w_{Li} \log(\text{OR}_i) / \sum w_{Li}$ . For the risk differences,  $w = w_{Ri}$ ,  $Y_i = \text{RD}_i$ , and  $Y_{\text{fix}} = \sum w_{Ri} \text{RD}_i / \sum w_{Ri}$ . Under the null hypothesis of a common treatment effect among trials, these  $Q$ -statistics follow a  $\chi^2$  distribution with  $K - 1$  degrees of freedom [17].

The one-tailed  $p$ -value of the  $Q$ -statistic provides a convenient measure of heterogeneity, one that can be applied across different treatment effect metrics. For a given set of randomized trials, one may say that the trial risk differences ‘display more heterogeneity’ (or ‘are more heterogeneous’) than the corresponding odds ratios, if the  $p$ -value for the risk difference  $Q$ -statistic is less than the  $p$ -value for the odds ratio  $Q$ -statistic. Similarly, following common practice, one may label a collection of trial odds ratios or risk differences as ‘heterogeneous’ when the corresponding  $Q$ -statistic  $p$ -value is below a nominal cut-off, usually 0.05 or 0.10 [17].

#### 2.4. Computer software

All computations used to calculate and compare the statistics presented are easily programmed. We used the computer program Meta-Analyst (version 0.989, © J. Lau, New England Medical Center, Boston, 1996; available upon request) to derive these summary statistics for each meta-analysis. For subsequent comparisons of meta-analysis statistics, we used S-plus (version 3.3 for Windows, © MathSoft, Seattle, 1995).

### 3. RESULTS

#### 3.1. Description of included meta-analyses

A total of 125 meta-analyses were included in this study: 80 were from medical journals, and 45 were from the CCPC (see Schmid *et al.* [18] for a list of 115 of these meta-analyses, to which 10 additional meta-analyses were added in a 1996 updated search [19–28]). As shown in Table II, the meta-analyses from the medical journals included more trials than did those in the CCPC. The journal-published meta-analyses therefore had more patients than those from the CCPC, although the average trial size within meta-analyses was comparable.

#### 3.2. Heterogeneity of meta-analyses

Figure 1 provides a plot of the  $p$ -value of the risk difference  $Q$ -statistic against the  $p$ -value of the odds ratio  $Q$ -statistic, for each of the 125 meta-analyses. As noted, the  $p$ -value of the  $Q$ -statistic measures heterogeneity among observed odds ratios or risk differences. We explored several scales for graphically displaying these  $p$ -values, including linear, logarithmic and square root scales. We display  $p$ -values in Figure 1 (and subsequently in Figures 3–6) using a scale proportional to the *fourth root* of the  $p$ -value, because this scale best displays these  $p$ -values over a broad range from 0 to 1, and it centres  $p$ -values lying between 0.05 and 0.10.

Table II. Comparison of meta-analyses published in medical journals with those in Cochrane Collaboration Pregnancy and Childbirth database (CCPC).

	All meta-analyses ( <i>N</i> = 125)	Journal-published meta-analyses ( <i>N</i> = 80)	CCPC meta-analyses ( <i>N</i> = 45)	<i>P</i> -value journal versus CCPC meta-analyses
Number (%) of meta-analyses with:				0.0002
6–10 trials	66 (53)	33 (41)	33 (73)	
11–15 trials	28 (22)	20 (25)	8 (18)	
≥ 16 trials	31 (25)	27 (34)	4 (9)	
Overall number of subjects in meta-analysis, median (interquartile range)	2485 (1353–7105)	4108 (1542–11,220)	1835 (1193–2940)	0.005
Average trial size, median (interquartile range)	260 (137–583)	263 (152–624)	177 (128–420)	0.16

The Wilcoxon rank sum test was used to compare distributions of these continuous variables between the two groups of meta-analyses.

The diagonal line in Figure 1 indicates where the  $Q$ -statistic  $p$ -value for the risk difference equals that for the odds ratio. It is apparent that the risk differences usually displayed more heterogeneity than the odds ratios; for 107 (86 per cent) of the meta-analyses, the  $Q$ -statistic  $p$ -value for the risk differences was less than that for the odds ratios (sign test,  $p < 0.0001$ ).

Regardless of which  $Q$ -statistic  $p$ -value we used as a cut-off to identify heterogeneity, more meta-analyses had heterogeneous risk differences than heterogeneous odds ratios. For the risk difference metric, and  $p$ -value cut-offs of 0.05, 0.10 (horizontal line, Figure 1) and 0.20, the numbers of meta-analyses judged heterogeneous were 56, 59 and 71, respectively. Using the odds ratio metric, the corresponding numbers of meta-analyses judged heterogeneous were 31, 44 (demarcated by the vertical line, Figure 1) and 54, respectively.

### 3.3. Comparison of fixed and random effects statistics

For the risk difference metric, we directly compared fixed and random effects statistics for each meta-analysis (Figure 2). Part (a) compares the point estimates, part (b) compares the standard errors and part (c) compared the  $Z$ -statistics, which are ratios of the point estimates to their standard errors. For 33 meta-analyses (26 per cent of the total), the fixed and random effects point estimates, standard errors and  $Z$ -statistics were all equal; these are the meta-analyses for which the value of the  $Q$ -statistic was less than  $K - 1$  (see Section 2.3.1). This equality is commonly interpreted as indicating that no appreciable heterogeneity is present and that fixed effects estimates are acceptable.

For 71 of the remaining 92 meta-analyses (77 per cent), the random effects risk difference estimate was more conservative (had a smaller  $Z$ -statistic) than the fixed effects estimate (Figure 2(c)). This was the case largely because random effects standard errors were larger than the corresponding fixed effects standard errors (Figure 2(b)); in fact, 75 of these 92 meta-analyses

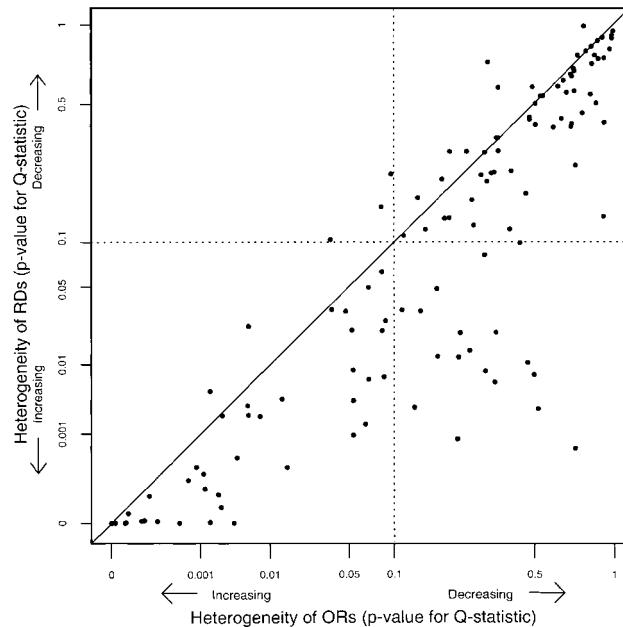


Figure 1.  $Q$ -statistic  $p$ -values (one-tailed) for trial odds ratios and risk differences are compared, for each of the 125 meta-analyses.  $P$ -values are plotted on a fourth-root scale (see Section 3.2). The diagonal line indicates equality, while the horizontal and vertical dashed lines indicate  $Q$ -statistic  $p$ -values of 0.10. Arrows indicate directions of increasing or decreasing heterogeneity.

(82 per cent) actually had a random effects point estimate more extreme (further from the null value of 0) than their fixed effects estimate (Figure 2(a)).

Similarly, for the odds ratio metric, we compared fixed and random effects statistics from each meta-analysis (Figures 2(d)–(f)). Because the random effects odds ratio statistic does not reduce to the Mantel–Haenszel odds ratio statistic when  $Q < K - 1$ , there was only one meta-analysis where the fixed and random effects summary odds ratios were equal (to four significant figures). Among the remaining 124 meta-analyses, neither method favoured more extreme summary odds ratios (the random effects estimates were closer to 1 for 63 meta-analyses,  $p = 0.93$  by the sign test). However, because the random effects standard errors were greater than the fixed effects standard errors in 119 meta-analyses (95 per cent), the random effects  $Z$ -statistics were usually less significant than the fixed effects  $Z$ -statistics (106, or 85 per cent, of meta-analyses, Figure 2(f); sign test,  $p < 0.0001$ ).

#### 3.4. Comparison of summary odds ratios and risk differences

We were interested in exploring whether the conclusions obtained by meta-analysis depend on whether one analyzes odds ratios or risk differences of the included trials. Although odds ratios and risk differences are not directly comparable, one can assess whether the summary estimates derived for these measures agree on the direction of treatment effect and the statistical significance of this effect.

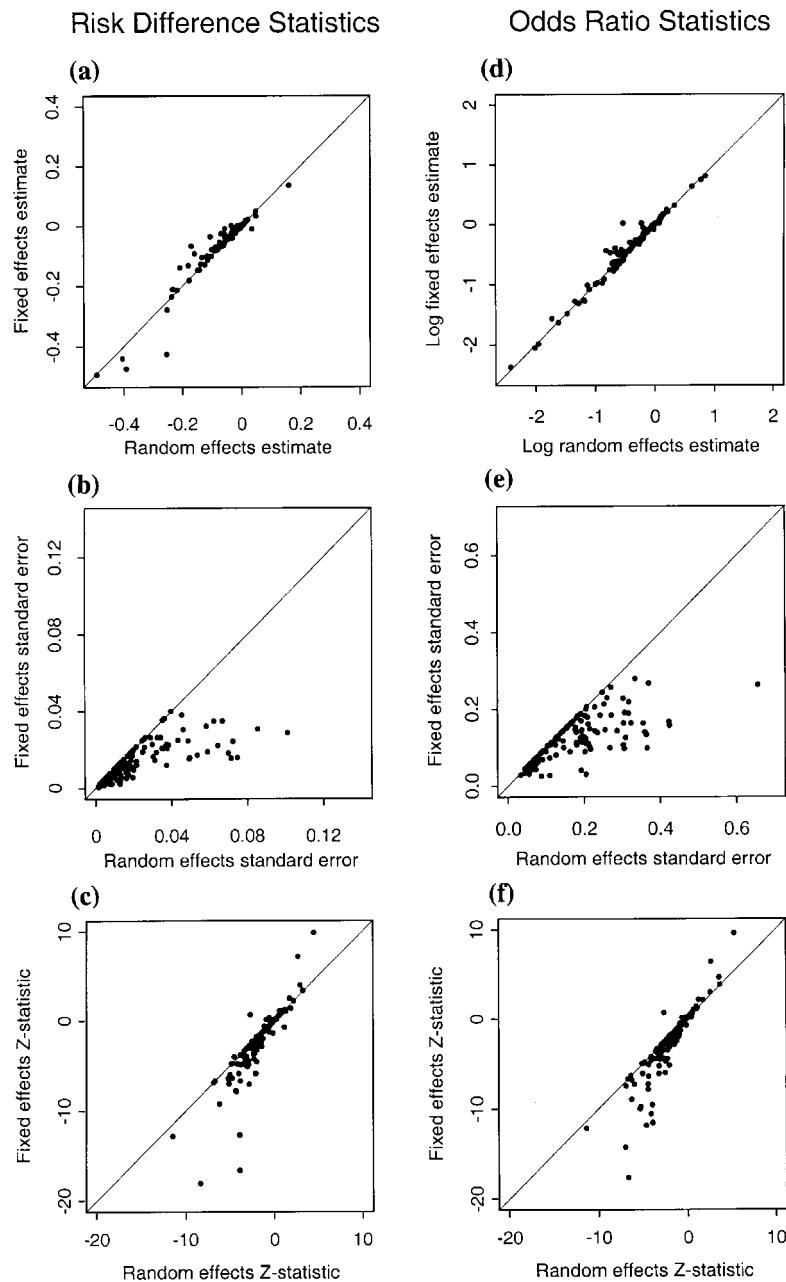


Figure 2. Fixed and random effects statistics are compared for each meta-analysis ( $N = 125$ ). For the risk difference metric, (a) compares fixed and random effects summary estimates, (b) compares the standard errors for these estimates, and (c) compares the Z-statistics for these estimates. Similarly, for the odds ratio, (d) compares fixed and random effects summary estimates (log-transformed), (e) compares standard errors for these log-transformed odds ratio estimates, and (f) compares their Z-statistics. In each part, the diagonal line indicates equality for the compared statistics.

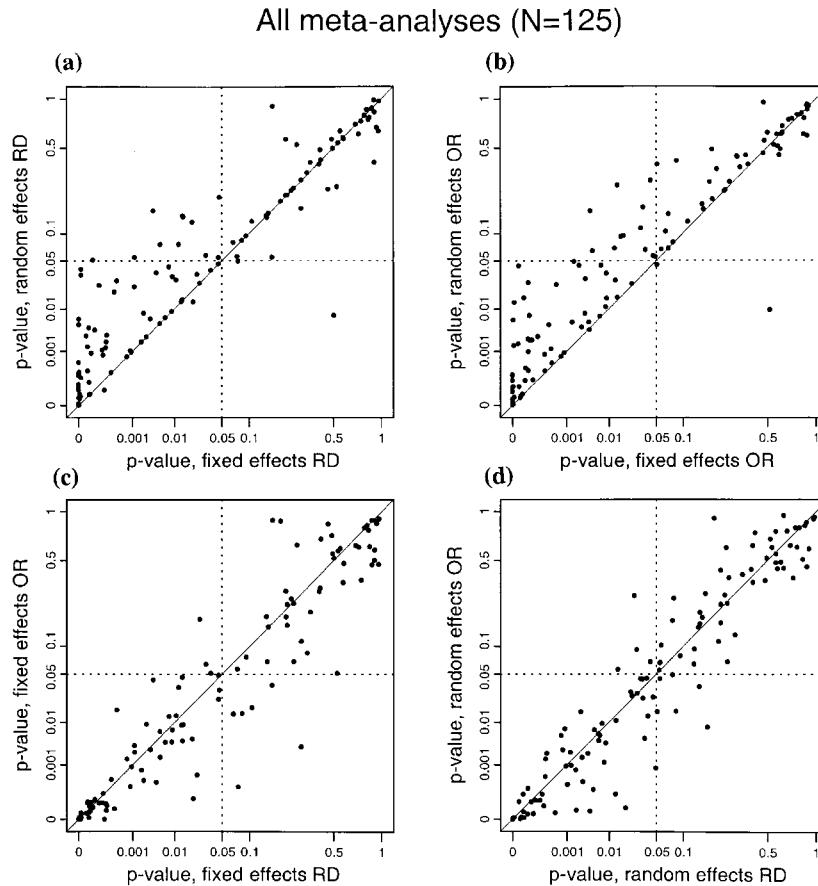


Figure 3. For each meta-analysis ( $N = 125$ ), the parts in this figure display pairwise comparisons of significance levels reached by four summary statistics: (a) random versus fixed effects risk difference; (b) random versus fixed effects odds ratio; (c) fixed effects odds ratio versus fixed effects risk difference; (d) random effects odds ratio versus random effects risk difference.  $P$ -values, which are two-sided, are displayed on a fourth root scale. Horizontal and vertical dashed lines indicate  $p$ -values of 0.05.

Of importance, we found no meta-analysis for which the summary risk difference and summary odds ratio indicated treatment effects that were each statistically significant but directed in opposite directions. This provides some comfort that, at least qualitatively, conclusions from meta-analyses are robust to changes of metric. As a result, we focused on the level of significance for the different summary estimates.

Figure 3 compares the significance levels (two-sided  $p$ -values) for four summary estimates: the fixed effects risk difference; fixed effects odds ratio; random effects risk difference, and random effects odds ratio. Figure parts comparing fixed and random effects risk differences (part (a)), and fixed and random effects odds ratios (part (b)), complement the  $Z$ -statistics shown in Figures 2(c) and (f); they illustrate that random effects estimates were often less significant than fixed effects estimates.

The fixed effects odds ratios were often more significant than the fixed effects risk differences (Figure 3(c)). However, these fixed effects methods were not always appropriate, since for 62 meta-analyses (50 per cent) the odds ratios and/or risk differences were heterogeneous ( $Q$ -statistic  $p$ -value  $< 0.10$ ). For the 125 meta-analyses overall, there was no tendency for the random effects odds ratio to be more significant than the random effects risk difference (sign test,  $p = 0.32$ ), though there was a reasonable amount of scatter around the line of equality for the respective  $p$ -values (Figure 3(d)).

Next, we divided the 125 meta-analyses into four subgroups depending on whether the odds ratios or risk differences were heterogeneous, defined for each metric as a  $Q$ -statistic  $p$ -value less than 0.10. These subgroups are described in Sections 3.4.1–3.4.4.

**3.4.1. 'Homogeneous' meta-analyses.** We describe as 'homogeneous' the subset of 63 meta-analyses (50 per cent) in which neither the odds ratios nor the risk differences were heterogeneous. Even within this subset, the risk differences displayed more heterogeneity than the odds ratios (sign test for  $Q$ -statistic  $p$ -values,  $p < 0.0001$ ).

As expected, for homogeneous meta-analyses, the significance levels of the random effects estimates were very close to those of the corresponding fixed effects estimates. With the risk difference metric, the significance levels of the fixed and random effects estimates were equal for 31 of these meta-analyses (49 per cent); for 22 of the remaining 32 meta-analyses the fixed effects estimate was more significant (Figure 4(a); sign test,  $p = 0.05$ ). For the odds ratio metric, the significance of fixed effects estimates generally agreed with that of random effects estimates (Figure 4(b)), although the fixed effects estimate was more significant for 47 meta-analyses (75 per cent,  $p < 0.0001$  sign test).

For almost all homogeneous meta-analyses, the significance of the fixed effects odds ratio agreed closely with that of the fixed effects risk difference (Figure 4(c); sign test,  $p = 0.90$ ). The random effects estimates for these meta-analyses, similar to the fixed effects estimates, also tended to agree (Figure 4(d)).

Of interest, one meta-analysis (indicated with an open symbol in Figure 4(c)) yielded a fixed effects odds ratio with much greater significance ( $p$ -value  $2.2 \times 10^{-5}$ ) than the corresponding fixed effects risk difference ( $p$ -value 0.02). Table III summarizes relevant aspects of this meta-analysis, which examined ACE-inhibitors in the treatment of heart failure [29]. Strikingly, due to the SOLVD trial's comparatively large size (2569 subjects, 36 per cent of total subjects in the meta-analysis) and high event rates  $p_T$  and  $p_C$ , its odds ratio of 0.80 carried 70 per cent of the weight in determining the fixed effects odds ratio. The Riegger and Uprichard trials, both much smaller and finding few events, carried very little weight. On the other hand, the fixed effects risk difference was pulled close to the null value by the small risk differences in the Riegger and Uprichard trials, which were precisely estimated and carried substantial weight. This meta-analysis therefore represents a case in which conclusions regarding treatment efficacy might be strongly affected by the metric used, in large part because fixed effects weights for odds ratios differ dramatically from those for risk differences.

These results are what one would expect from the earlier discussion of Table I, where we noted that studies with low event rates  $p_T$  and  $p_C$  receive large weight using the risk difference metric and studies with high event rates receive large weight with the odds ratio metric. Owing to differences in the event rates in trials of the ACE-inhibitor meta-analysis, the relatively large SOLVD trial contributes greatly to the summary odds ratio estimate but little to the risk difference estimate.

## Meta-analyses with neither OR nor RD heterogeneous (N=63)

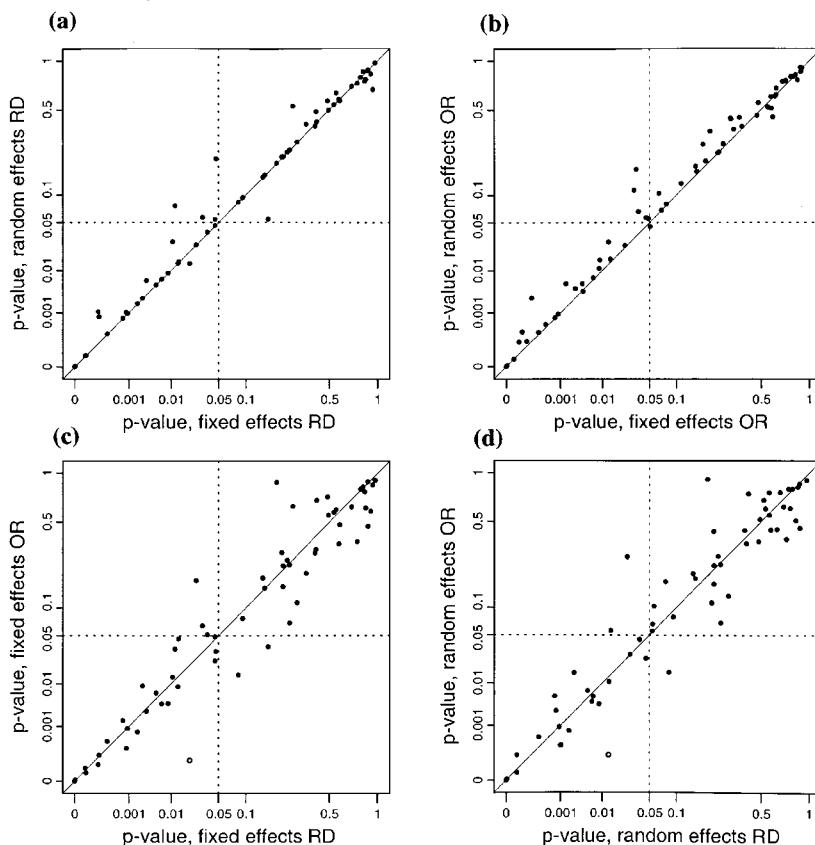


Figure 4. Pairwise comparisons of the significance levels for four summary statistics are presented, as in Figure 3, for the subset of 63 'homogeneous' meta-analyses. Homogeneous meta-analyses are those for which both odds ratio and risk difference  $Q$ -statistics are non-significant ( $p \geq 0.10$ ). In parts (c) and (d), the meta-analysis by Garg *et al.* [29] is represented by an open symbol (see Section 3.4.1).

3.4.2. *Meta-analyses in which both risk differences and odds ratios were heterogeneous.* There were 41 meta-analyses in which both the odds ratios and risk differences were heterogeneous. For 38 of these, the risk differences were more heterogeneous than the odds ratios (sign test,  $p < 0.0001$ ).

For both odds ratio and risk difference metrics, the random effects estimates were less significant than the corresponding fixed effects estimates (Figures 5(a) and (b); sign test,  $p < 0.0001$  for each comparison). Of note, however, for one of these meta-analyses (open symbol, Figures 5(a) and (b)), the random effects odds ratio and risk difference were both substantially more significant than the respective fixed effects estimates. This meta-analysis, which examined mortality with magnesium treatment following myocardial infarction, has been described in several reports [30–32].

Table IV provides a summary of odds ratio data for the magnesium therapy meta-analysis. The ISIS-4 trial, largest by far in the meta-analysis with 94 per cent of the total patients, found no

Table III. Description of meta-analysis of ACE-inhibitors in heart failure.

Trial	Number in treatment arm	$p_T$	Number in control arm	$p_C$	Odds ratio	Weight for OR in fixed effects OR*	Risk difference	Weight for RD in fixed effects RD*
SOLVD	1285	0.352	1284	0.405	0.80	70.1%	-0.05	5.3%
Riegger	169	0.000	56	0.000	0.33	0.1%	0	11.3%
Uprichard	139	0.014	47	0.000	1.73	0.2%	0.01	18.8%
Other 29 trials	2277	0.069	1848	0.108	0.64 <sup>†</sup>	29.6%	-0.02 <sup>†</sup>	64.6%
Summary statistics for meta-analysis	Number of trials: 32	Q-stat OR 20.8 ( $p = 0.92$ )	Q-stat RD 31.8 ( $p = 0.14$ )		Fixed effects OR 0.75 ( $p = 2.2 \times 10^{-5}$ )		Fixed effects RD -0.01 ( $p = 0.02$ )	

Data derived from Garg *et al.* [29].

Abbreviations:  $p_T$ , proportion with outcome in treatment arm;  $p_C$ , proportion with outcome in control arm; OR, odds ratio; RD, risk difference; Q-stat, Q-statistic.

\* These weights are expressed as the proportion of the total weights for all trials in the meta-analysis that was assigned to the given trial.

† These estimates are fixed effects estimates summarized for the 29 trials in this row.

## Meta-analyses with both OR and RD heterogeneous (N=41)

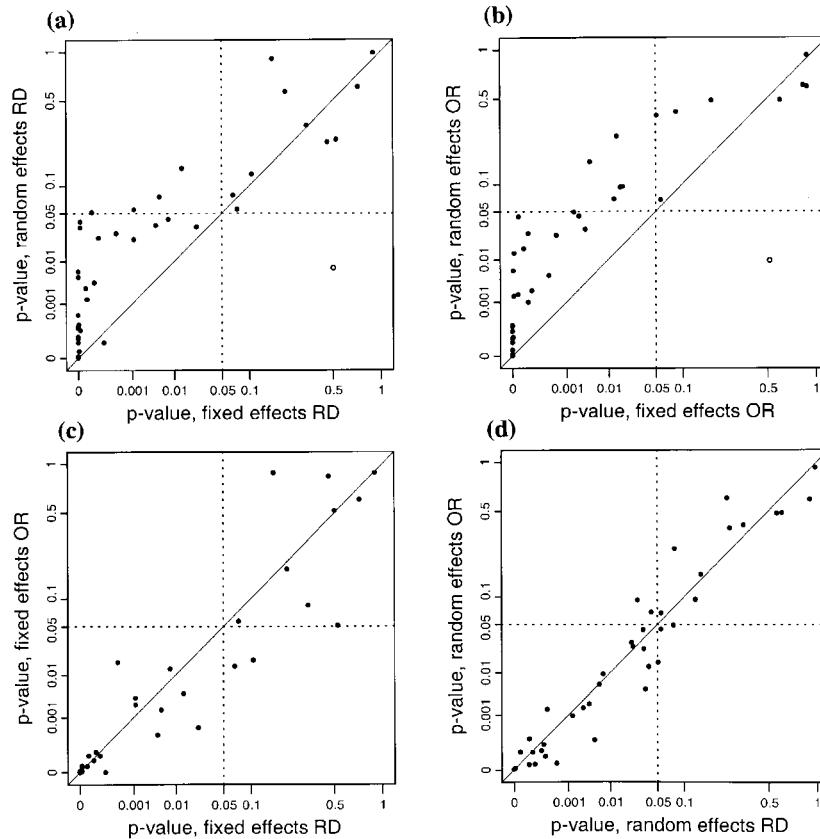


Figure 5. Pairwise comparisons of the significance levels for four summary statistics are presented, as in Figure 3, for the subset of 41 meta-analyses in which both odds ratios and risk differences are heterogeneous ( $Q$ -statistic  $p < 0.10$ ). In parts (a) and (b), the meta-analysis of magnesium therapy is represented by an open symbol (see Section 3.4.2).

benefit to treatment, and its odds ratio of 1.06 pulled the fixed effect estimate toward the null value. However, for the random effects odds ratio estimate, substantial heterogeneity among the 11 trial odds ratios led to a variance component  $D$  substantially greater than zero. This caused the relative weights to be more equal under the random effects model than under the fixed effects model, as was discussed in Section 2.3.1. The random effects summary estimate was less than the null value, due to the influence of smaller trials that found a treatment benefit. For this meta-analysis, the random effects estimate was statistically significant, even with its larger standard error. A similar analysis applies for the risk difference estimates in this meta-analysis.

As shown in Figure 5(d), the random effects odds ratio and risk difference estimates agreed moderately well in statistical significance (sign test,  $p = 0.35$ ). For completeness, part (c) compares fixed effects estimates, although they are not appropriate in the presence of heterogeneity.

Table IV. Description of meta-analysis of magnesium following myocardial infarction.

Trial	Total in treatment arm	$p_r$	Total in control arm	$p_c$	Odds ratio	Weight for OR in fixed effects OR*	Weight for OR in random effects OR*
ISIS-4	29011	0.076	29039	0.072	1.06	93.8%	24.9%
LIMIT-2	1159	0.077	1157	0.102	0.74	4.4%	22.2%
Rasmussen	135	0.067	135	0.170	0.35	0.5%	12.4%
Other eight trials	687	0.035	664	0.080	0.41 <sup>†</sup>	1.3%	40.5%
Summary statistics for meta-analysis	Number of trials: 11	$Q$ -stat OR 31.2 ( $p = 0.0006$ )			Fixed effects OR 1.02 ( $p = 0.52$ )		Random effects OR 0.59 ( $p = 0.009$ )

Data derived from Teo *et al.* [31] and two updates [30, 32].

Abbreviations: OR, odds ratio;  $Q$ -stat,  $Q$ -statistic.

\* These weights are expressed as the proportion of the total weights for all trials in the meta-analysis that was assigned to the given trial.

<sup>†</sup> This odds ratio is the random effects odds ratio summarized for the eight trials in this row.

## Meta-analyses with only RD heterogeneous (N=18)

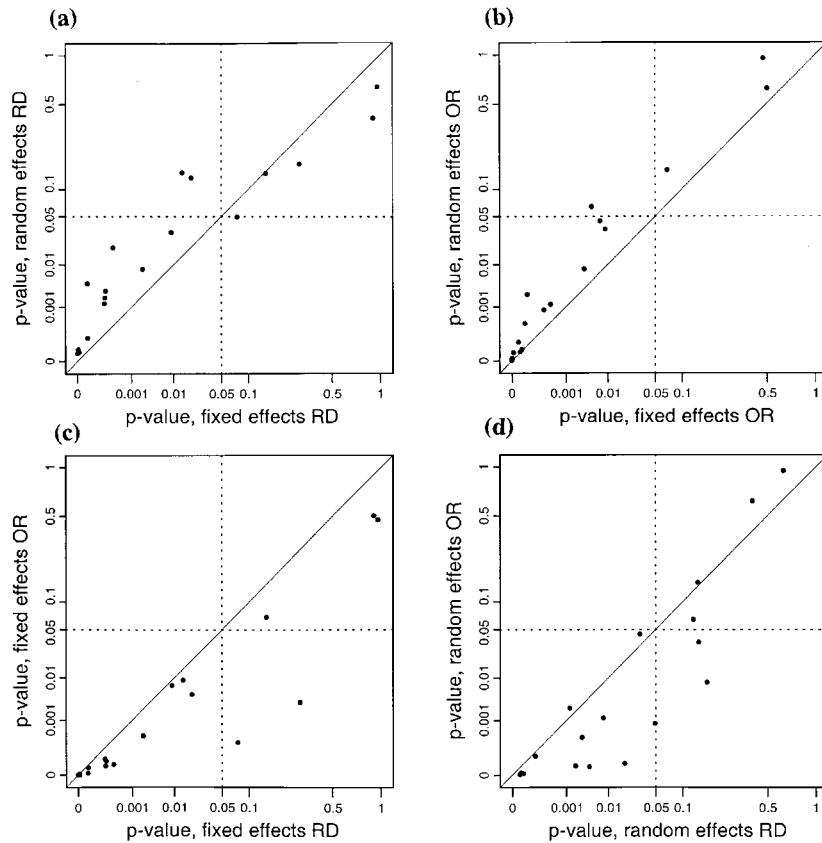


Figure 6. Pairwise comparisons of the significance levels for four summary statistics are presented, as in Figure 3, for the subset of 18 meta-analyses in which only risk differences are heterogeneous ( $Q$ -statistic  $p < 0.10$ ).

**3.4.3. Meta-analyses in which only the risk differences were heterogeneous.** There were 18 meta-analyses in which only the risk differences were judged heterogeneous. With the risk difference metric, the random effects estimate was less significant than the fixed effects estimate for 13 of these meta-analyses (Figure 6(a); sign test,  $p = 0.10$ ).

For the odds ratio metric, Figure 6(b) compares the significance of the fixed and random effects estimates. The random effects odds ratio was less significant than the fixed effects odds ratio for all 18 of these meta-analyses (sign test,  $p < 0.0001$ ). This is explained by the *relative* heterogeneity of the odds ratios in these meta-analysis; even though the odds ratios were not significantly heterogeneous in any meta-analysis (each had  $Q$ -statistic  $p \geq 0.10$ ), the odds ratios were still heterogeneous enough for fixed and random effects estimates to differ.

As shown in Figure 6(d), the random effects odds ratio was more significant than the random effects risk difference for 13 meta-analyses (72 per cent; sign test,  $p = 0.10$ ). Again the significance

levels of the fixed effects estimates are compared graphically in Figure 6(c), although it is not appropriate to use a fixed effects method for the risk differences.

*3.4.4. Meta-analyses in which only the odds ratios were heterogeneous.* There were only three meta-analyses in which the odds ratios were heterogeneous while the risk differences were not heterogeneous; these meta-analyses lie in the upper left portion of Figure 1. These meta-analyses are a subset of the 18 meta-analyses that lie above the diagonal line in that figure, for each of which the odds ratios were more heterogeneous than the risk differences. As can be seen from Figure 1, the  $Q$ -statistic  $p$ -values for the odds ratios and risk differences are not dramatically different for these three meta-analyses, and the meta-analyses are identified solely by the arbitrary cut-off point of 0.10 for the  $Q$ -statistic  $p$ -value. We therefore do not display separate plots of the  $p$ -values for the summary estimates for these three meta-analyses.

*3.4.5. Nominal significance obtained for each method of effect summary.* Because a  $p$ -value cut-off of 0.05 is frequently used to determine statistical significance of treatment effect estimates, this cut-off has been marked in Figures 3–6. Table V displays the corresponding number of meta-analyses judged significant at the 0.05 level, for each of the summary methods.

Using either random effects summary measure, more than half of meta-analyses in our sample were nominally significant. This observation suggests that meta-analyses may often alert clinicians to treatment differences or confirm the existence of important clinical effects. For either random effects measure, the same proportion of meta-analyses were significant for journal-published and CCPC meta-analyses (Table V).

Several explanations may be advanced for this large proportion of significant meta-analyses. It is likely that treatments are only evaluated in clinical trials when prior evidence indicates that they are likely to be effective. This selection would lead to a preponderance of published trials with treatment benefits, with the result that large meta-analyses of these trials (presumably with substantial power) would frequently document benefit. It is also possible that investigators only chose to submit findings for publication when statistically significant treatment effects were demonstrated (the file-drawer phenomenon [33]). Alternatively, it may be true that investigators who perform and publish meta-analyses wait for enough trials to ensure that they have sufficient power to uncover a treatment benefit. Just as single trials may often remain unpublished if they are small and non-significant, so too may meta-analyses fail to be published as not newsworthy if they are inconclusive. This explanation is not supported, however, by our finding that meta-analyses from the more inclusive CCPC were as likely to be significant as those from the peer-reviewed journals. Finally, the observation that meta-analyses frequently reach statistical significance may be related in part to our minimum size requirements, in that collections of fewer than six trials were excluded from the present study.

An unexpected finding in Table V is that homogeneous meta-analyses less often produced statistically significant summary effect estimates than other meta-analyses, in which odds ratios or risk differences were heterogeneous. For example, 37 per cent of homogeneous meta-analyses produced a significant random effects odds ratio, compared with 71 per cent of non-homogeneous meta-analyses ( $\chi^2$  test,  $p = 0.0002$ ). A tentative explanation for this observation may be outlined as follows. Because separate trials are unlikely to find the same treatment both substantially beneficial and substantially harmful, heterogeneity may usually arise only when some trials find a treatment effect in one direction while others find no treatment effect. The summary estimate derived from these trials may then be pulled away from the null value, toward statistical

Table V. Meta-analysis reaching statistical significance.

Subgroup of meta-analyses	Number (%) of meta-analyses that were statistically significant, using the specified summary method*			
	Fixed effects odds ratio	Random effects odds ratio	Fixed effects risk difference	Random effects risk difference
All meta-analyses ( $N = 125$ )	77 (62)	67 (54)	74 (59)	65 (52)
Journal-published meta-analyses ( $N = 80$ )	51 (64)	44 (55)	48 (60)	43 (54)
CCPC meta-analyses ( $N = 45$ )	26 (58)	23 (51)	26 (58)	22 (49)
Homogeneous meta-analyses <sup>†</sup> ( $N = 63$ )	27 (43)	23 (37)	28 (44)	24 (38)
All other meta-analyses ( $N = 62$ )	50 (81)	44 (71)	46 (74)	41 (66)
Subgroups in which meta-analyses heterogeneous <sup>‡</sup> for:				
Both risk differences and odds ratios ( $N = 41$ )	32 (78)	28 (68)	30 (73)	27 (66)
Only risk differences ( $N = 18$ )	15 (83)	14 (78)	13 (72)	12 (67)
Only odds ratios ( $N = 3$ )	3 (100)	2 (67)	3 (100)	2 (67)

CCPC: Cochrane Collaboration Pregnancy and Childbirth Database.

\* Summary estimates are judged significant if the  $p$ -value (two-sided) is less than 0.05.

<sup>†</sup> Homogeneous meta-analyses are meta-analyses for which the  $Q$ -statistic  $p \geq 0.10$  for both odds ratios and risk differences.

<sup>‡</sup> Within meta-analyses, heterogeneity is defined for the odds ratios and for the risk differences as  $p < 0.10$  for the corresponding  $Q$ -statistic.

significance. In contrast, homogeneous meta-analyses may examine trials whose effect estimates are clustered around the null value (otherwise no meta-analysis would be necessary), and summary estimates may tend not to be significant. Another possible explanation for heterogeneous meta-analyses being more likely to find a significant treatment effect may be that meta-analyses with adequate power to detect heterogeneity may tend to be the same meta-analyses with adequate power for finding a treatment effect. Further study is needed to confirm this observation and evaluate these explanations.

#### 4. DISCUSSION

To our knowledge, this empirical study represents the largest systematic examination of meta-analyses. We studied a sample of 125 meta-analyses, representative of those performed by clinical investigators, to determine how characteristics of the collections of trials and statistical summary methods influence the conclusions drawn from meta-analyses. We compared two separate ways of measuring treatment effect (odds ratio and risk difference), looked at agreement between fixed and random summary estimates, and examined the prevalence of heterogeneity and its impact on effect estimates. Although our study is primarily descriptive in nature, several findings shed light on important theoretical issues and have immediate practical implications.

#### 4.1. Heterogeneity

Using the  $Q$ -statistic  $p$ -value to measure heterogeneity, we found that the trial risk differences displayed more heterogeneity than trial odds ratios. For 86 per cent of the meta-analyses, the risk differences were more heterogeneous than the trial odds ratios, and differences in heterogeneity were often substantial (Figure 1). For 18 meta-analyses (14 per cent) the risk differences were judged heterogeneous ( $Q$ -statistic  $p < 0.10$ ) when the odds ratios were not, whereas for only three meta-analyses (2 per cent) were the odds ratios heterogeneous when the risk differences were not.

Two studies have previously examined the heterogeneity of individual trial effect estimates within published meta-analyses; in contrast to our findings, they documented similar heterogeneity measures for risk differences and odds ratios [14, 34]. However, because these previous studies were quite small (nine meta-analyses in the study by DerSimonian and Laird, 22 in the study by Berlin *et al.*), they may have lacked sufficient power to detect differences in heterogeneity between risk differences and odds ratios. Furthermore, both previous studies used convenience samples which may not have been representative of meta-analyses found in clinical research. Our much larger study sample was drawn systematically from medical journals and the Cochrane Collaboration Pregnancy and Childbirth database [10], to reflect a broad range of meta-analyses. Also, some meta-analyses in the previous studies were smaller than those allowed in the present study. For instance, Berlin *et al.* [34] included nine meta-analyses with fewer than six trials, meta-analyses that we would have excluded.

Trial risk differences may display more heterogeneity than odds ratios because they are more closely correlated with the proportion of subjects in the control group who develop the outcome of interest (control rate). The control rate can be thought of as a measure of the underlying risk for subjects in the trial, although it is affected not just by the health of the trial population but by trial characteristics, such as follow-up time and surveillance intensity [18]. If the treatment is more effective than the control, the risk difference will tend to be negative. In such a case, as the control rate approaches zero, so must the risk difference, because the rate in the treatment arm cannot be negative. Therefore, although risk differences can remain constant in a narrowly defined context, wide variations in trial populations or trial characteristics will affect the control rate and thus the risk difference. Odds ratios are not constrained by control rates in the same way as risk differences, and empirically odds ratios are less correlated with trial control rates than are risk differences [18]. How much of the heterogeneity of treatment effects is accounted for by variation in trial control rates is an area of active investigation.

#### 4.2. Comparison of random and fixed effects summary estimates

Since heterogeneity is incorporated directly into random effects summary estimates and their standard errors, it is not surprising that random effects estimates sometimes differed from corresponding fixed effects estimates. The most obvious effect of heterogeneity was to increase the standard errors of the random effects estimates (Figures 2(b) and (e)). Because the random effects estimates themselves did not differ much from the corresponding fixed effects estimates (Figures 2(a) and (d)), the overall effect of heterogeneity was to make most random effects estimates less significant than the corresponding fixed effects estimates, for both the odds ratio and risk difference metrics.

For one meta-analysis in this study, that of magnesium treatment following myocardial infarction, the random effects summary estimates indicated a larger treatment effect and were more significant than the corresponding fixed effects estimates (Figure 5 and Table IV). This

collection of trials generated controversy about the comparability of large trials and meta-analyses [30, 35], because a meta-analysis of the small trials found a benefit to magnesium treatment [31], whereas the mega-trial ISIS-4 subsequently found no benefit. Table IV shows the substantial heterogeneity among trial results that led to this controversy; as measured by  $Q$ -statistic  $p$ -values for odds ratios and risk differences, this meta-analysis was among the most heterogeneous meta-analyses in our study. The magnesium meta-analysis therefore serves as a warning to investigators to examine heterogeneous collections of trial results carefully and, if a summary effect measure is to be derived, to pay attention to weights assigned to individual trial results. Nonetheless, the behaviour of estimates in the magnesium meta-analysis is unusual, in that most heterogeneous meta-analyses that we examined demonstrated random effects estimates similar in magnitude to and less significant than fixed effects estimates.

#### 4.3. Comparison of summary odds ratios and risk differences

For most meta-analyses in which neither odds ratios nor risk differences were heterogeneous, fixed effects odds ratios and fixed effects risk differences provided similar levels of significance (Figure 4(c)). Furthermore, when both odds ratios and risk differences were heterogeneous, random effects odds ratio and risk difference summary estimates tended to demonstrate similar levels of statistical significance (Figure 5(d)). This finding that summary odds ratios and risk differences often agree with respect to the degree of statistical significance is reassuring and suggests that the choice of metric used to measure and summarize the treatment effect is not crucial. For the single meta-analysis that we identified in which the statistical significance of risk difference and odds ratio summary estimates greatly differed, there were clear differences among trial sizes and event rates (Table III). As demonstrated in Table I, differences in event rates can lead to very different relative weights being applied by different metrics to the same study.

#### 4.4. Implications for the conduct of meta-analyses

The risk difference may not be the most appropriate metric to use in meta-analysis, because risk differences may be substantially heterogeneous among trials. Furthermore, the risk difference metric tends to give greatest weight to trials with low event rates (Table I). This is counterintuitive, because trials with low event rates would seem to offer little information about treatment effects [36]. Also, in general, the asymptotic assumptions upon which standard meta-analytic estimates are based are not supported by trials with few events. How commonly meta-analyses include trials with low event rates, and how frequently this affects summary estimates, cannot be determined from this study, because we specifically excluded some meta-analyses with low event-rate trials.

Clinical and public policy decisions are often based on absolute measures of treatment effect, such as the risk difference, or its multiplicative inverse, the number-needed-to-treat. Our results suggest that the odds ratio is more likely than the risk difference to remain constant across populations. Therefore, when calculating the absolute treatment benefit expected for a patient, it may be optimal first to perform a meta-analysis of available trial data on the odds ratio scale and then apply the summary odds ratio to the patient's expected risk [9].

Investigators should assess heterogeneity of trial results before deriving summary estimates of treatment effect [3–5]. Heterogeneity, even for trial odds ratios, is common among collections of trials that clinical investigators examine. While fixed effects odds ratio estimates tend to agree in magnitude with random effects estimates, when heterogeneity is present fixed effects standard

errors often suggest inappropriate precision. Given the low power of the  $Q$ -statistic [37], some investigators believe it is most appropriate to assume that systematic differences among trials are always present, even when the  $Q$ -statistic is non-significant, and to use a random effects summary of treatment effect [38]. In any event, when trial results are markedly heterogeneous, a clustering procedure or regression approach will usually be more appropriate than any summary method that estimates a single 'average' effect [3, 4, 39].

#### 4.5. Directions for future research

There are several avenues for further research. Some issues, such as the dependence of summary estimates on weights assigned to individual trials, may best be approached through detailed simulation studies. One might also assess through simulation the magnitude of bias of fixed and random effects odds ratio estimators. Although we have presented several lines of evidence that support use of summary odds ratio estimates in meta-analysis, the bias of these estimates may be substantial when the trials are small, because the estimates are linear combinations of observed trial odds ratios or their logarithm [12].

It will be important to confirm our empirical findings in other collections of meta-analyses, such as those in other medical journals or in the Cochrane Library, a collection of meta-analyses submitted by investigators on a wide range of topics [40]. Additionally, it will be useful to evaluate other measures of treatment effect. An obvious choice is the risk ratio; our preliminary work suggests that risk ratios, like odds ratios, are usually less heterogeneous than risk differences. Examination of a measure with stabilized variance, the 'arcsin' metric [7], may also be informative.

#### ACKNOWLEDGEMENTS

This research report was developed under contract with the Agency for Healthcare Research and Quality (contract number 290970019). Eric Engels also received support from the Agency for Healthcare Research and Quality, grant number T32 HS00060. Christopher Schmid and Joseph Lau received support from the Agency for Healthcare Research and Quality, R01 HS08532 and HS10064, as well as from the National Center for Research Resources, R41 RR12416.

#### REFERENCES

1. Chalmers TC, Lau J. Meta-analytic stimulus for changes in clinical trials. *Statistical Methods in Medical Research* 1993; **2**:161–172.
2. Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *New England Journal of Medicine* 1992; **327**:248–254.
3. Greenland S. Invited commentary: A critical look at some popular meta-analytic methods. *American Journal of Epidemiology* 1994; **140**:290–296.
4. Lau J, Ioannidis JPA, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet* 1998; **351**:123–127.
5. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal* 1994; **309**:1351–1355.
6. Snedecor GW, Cochran WG. *Statistical Methods*. Iowa State University Press: Ames, Iowa, 1980; 287–292.
7. Emerson JD. Introduction to transformation. In *Fundamentals of Exploratory Analysis of Variance*, Hoaglin DC, Mosteller F, Tukey JW (eds). Wiley: New York, 1991.
8. Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *Journal of Clinical Epidemiology* 1994; **47**:881–889.
9. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *British Medical Journal* 1995; **310**:452–454.

10. Enkin MW, Keirse MJNC, Renfrew MJ, Neilson JP. Pregnancy and Childbirth Module: Cochrane Database of Systematic Reviews. *Cochrane Updates on Disk*. Oxford, 1994.
11. Cox DR, Snell EJ. *Analysis of Binary Data*. Chapman and Hall: New York, 1989; 32.
12. Agresti A. *Categorical Data Analysis*. Wiley: New York, 1990.
13. Robins J, Greenland S, Breslow NE. A general estimator for the variance of the Mantel-Haenszel odds ratio. *American Journal of Epidemiology* 1986; **124**:719-723.
14. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177-188.
15. Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *Journal of Clinical Epidemiology* 1991; **44**:127-139.
16. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine* 1991; **10**:1665-1677.
17. Hedges LV, Olkin I. *Statistical Methods for Meta-analysis*. Academic Press: New York, 1985.
18. 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**:1923-1942.
19. Early Breast Cancer Trialists Collaborative Group. Ovarian ablation in early breast cancer: overview of the randomised trials. *Lancet* 1996; **348**:1189-1196.
20. Childhood ALL Collaborative Group. Duration and intensity of maintenance chemotherapy in acute lymphoblastic leukaemia: overview of 42 trials involving 12000 randomised children. *Lancet* 1996; **347**:1783-1788.
21. Barza M, Ioannidis JP, Cappelleri JC, Lau J. Single or multiple daily doses of aminoglycosides: a meta-analysis. *British Medical Journal* 1996; **312**:338-345.
22. Collins R, MacMahon S, Flather M, Baigent C, Remvig L, Mortensen S, Appleby P, Godwin J, Yusuf S, Peto R. Clinical effects of anticoagulant therapy in suspected acute myocardial infarction: systematic overview of randomised trials. *British Medical Journal* 1996; **313**:652-659.
23. Gelber RD, Cole BF, Goldhirsch A, Rose C, Fisher B, Osborne CK, Boccardo F, Gray R, Gordon NH, Bengtsson NO, Sevela P. Adjuvant chemotherapy plus tamoxifen compared with tamoxifen alone for postmenopausal breast cancer: meta-analysis of quality-adjusted survival. *Lancet* 1996; **347**:1066-1071.
24. Ioannidis JP, Cappelleri JC, Skolnik PR, Lau J, Sacks HS. A meta-analysis of the relative efficacy and toxicity of *Pneumocystis carinii* prophylactic regimens. *Archives of Internal Medicine* 1996; **156**:177-188.
25. Linde K, Ramirez G, Mulrow CD, Pauls A, Weidenhammer W, Melchart D. St. John' wort for depression - an overview and meta-analysis of randomised clinical trials. *British Medical Journal* 1996; **313**:253-258.
26. Marchioli R, Marfisi RM, Carinci F, Tognoni G. Meta-analysis, clinical trials, and transferability of research results into practice. The case of cholesterol-lowering interventions in the secondary prevention of coronary heart disease. *Archives of Internal Medicine* 1996; **156**:1158-1172.
27. Oler A, Whooley MA, Oler J, Grady D. Adding heparin to aspirin reduces the incidence of myocardial infarction and death in patients with unstable angina. A meta-analysis. *Journal of the American Medical Association* 1996; **276**:811-815.
28. Roberts I, Kramer MS, Suissa S. Does home visiting prevent childhood injury? A systematic review of randomized controlled trials. *British Medical Journal* 1996; **312**:29-33.
29. Garg R, Yusuf S. Overview of randomized trials of angiotensin-converting enzyme inhibitors on mortality and morbidity in patients with heart failure. *Journal of the American Medical Association* 1995; **18**:1450-1456.
30. Borzak S, Ridker PM. Discordance between meta-analyses and large-scale randomized, controlled trials. Examples from the management of acute myocardial infarction. *Annals of Internal Medicine* 1995; **123**:873-877.
31. Teo KK, Yusuf S, Collins R, Held PH, Peto R. Effects of intravenous magnesium in suspected acute myocardial infarction: an overview of randomised controlled trials. *British Medical Journal* 1991; **303**:1499-1503.
32. Teo KK, Yusuf S. Role of magnesium in reducing mortality in acute myocardial infarction. A review of the evidence. *Drugs* 1993; **46**:347-359.
33. Gleser LJ, Olkin I. Models for estimating the number of unpublished studies. *Statistics in Medicine* 1996; **15**:2493-2507.
34. Berlin JA, Laird NM, Sacks HS, Chalmers TC. A comparison of statistical methods for combining event rates from clinical trials. *Statistics in Medicine* 1989; **8**:141-151.
35. LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New England Journal of Medicine*, 1997; **337**:536-542.
36. Engels EA, Lau J. Symptomless colonisation by *Clostridium difficile* and risk of diarrhoea. *Lancet* 1998; **351**:1733.
37. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine* 1998; **17**:841-856.
38. Hasselblad V, Mosteller F, Littenberg B, Chalmers TC, Hunink MG, Turner JA, Morton SC, Diehr P, Wong JB, Powe NR. A survey of current problems in meta-analysis. Discussion from the Agency for Health Care Policy and Research Inter-PORT work group on literature review/meta-analysis. *Medical Care* 1995; **33**:202-220.
39. Hedges LV, Olkin I. Clustering estimates of effect magnitude from independent studies. *Psychological Bulletin* 1983; **93**:563-573.
40. Cochrane Library [database]. Cochrane Collaboration; Oxford; 1999, updated quarterly.