The effects of clinical and statistical heterogeneity on the predictive values of results from meta-analyses

W. G. Melsen1, M. C. J. Bootsma1,2, M. M. Rovers3 and M. J. M. Bonten1,4
1) Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, 2) Department of Mathematics, Faculty of Science, Utrecht University, 3) Nijmegen Centre for Evidence Based Practice, Radboud University Nijmegen Medical Centre, Nijmegen and 4) Department of Medical Microbiology, University Medical Centre Utrecht, Utrecht, the Netherlands

Abstract

Variance between studies in a meta-analysis will exist. This heterogeneity may be of clinical, methodological or statistical origin. The last of these is quantified by the $I^2$-statistic. We investigated, using simulated studies, the accuracy of $I^2$ in the assessment of heterogeneity and the effects of heterogeneity on the predictive value of meta-analyses. The relevance of quantifying $I^2$ was determined according to the likely presence of heterogeneity between studies (low, high, or unknown) and the calculated $I^2$ (low or high). The findings were illustrated by published meta-analyses of selective digestive decontamination and weaning protocols. As expected, $I^2$ increases and the likelihood of drawing correct inferences from a meta-analysis decreases with increasing heterogeneity. With low levels of heterogeneity, $I^2$ does not appear to be predictive of the accuracy of the meta-analysis result. With high levels of heterogeneity, even meta-analyses with low $I^2$-values have low predictive values. Most commonly, the level of heterogeneity in a meta-analysis will be unknown. In these scenarios, $I^2$ determination may help to identify estimates with low predictive values (high $I^2$). In this situation, the results of a meta-analysis will be unreliable. With low $I^2$-values and unknown levels of heterogeneity, predictive values of pooled estimates may range extensively, and findings should be interpreted with caution. In conclusion, quantifying statistical heterogeneity through $I^2$-statistics is only helpful when the amount of clinical heterogeneity is unknown and $I^2$ is high. Objective methods to quantify the levels of clinical and methodological heterogeneity are urgently needed to allow reliable determination of the accuracy of meta-analyses.

Keywords: Heterogeneity, $I^2$-statistic, meta-analysis, systematic review

Article published online: 9 December 2013

Clin Microbiol Infect 2014; 20: 123–129

Introduction

The meta-analysis has become one of the most widely used methods to quantify the effects of medical interventions. In fact, in grading the evidence base of medical practice, a properly designed meta-analysis is considered to be equally as relevant as a large randomized controlled trial, as one of both is needed to reach so-called level I evidence [1]. As such, meta-analyses generally constitute the starting point, and frequently the most prominent component, of guidelines for clinical management. Furthermore, clinicians are increasingly using meta-analyses to remain up-to-date, and funding agencies frequently require such an analysis to justify further research. The number of published systematic reviews and meta-analyses has increased substantially in the last decade, including in the field of infectious disease medicine.
Ideally, a meta-analysis combines the results of several studies that are highly comparable in design, intervention, and patient population. The individual studies have similar trends in outcome, but lack sufficient statistical power for a definite conclusion to be drawn. However, in real life, meta-analyses frequently contain multiple, relatively small studies that differ in many respects (such as in dosing schedules, duration of follow-up, types of participants, and modes of treatment and diagnosis).

Naturally, studies brought together in a meta-analysis will differ, and this is also called ‘heterogeneity’. Generally, a distinction is made between clinical heterogeneity (differences in, for example, patient populations and treatment protocol), methodological heterogeneity (differences in study design and risk of bias), and statistical heterogeneity (larger differences in the outcome of the individual studies than could expected to result from chance alone, which may result from clinical or methodological heterogeneity).

Tests for statistical heterogeneity, such as Cochran’s Q-statistic and the $I^2$-statistic, are commonly used in meta-analyses to determine whether there are genuine differences underlying the results of the studies, or whether the variation in findings is compatible with chance alone. The most commonly used test is the $I^2$-statistic, which expresses the level of heterogeneity as a percentage, and can be compared across meta-analyses with different sizes and outcomes [2].

The appraisal of the similarity of studies with regard to clinical and methodological heterogeneity and the ultimate decision of whether to include (or exclude) a certain study in a meta-analysis are the responsibility of the meta-analysts. As there are no criteria with which to quantify clinical and methodological heterogeneity, this appraisal is subjective. Although the quantification of statistical heterogeneity seems to be more objective (e.g. by calculating the $I^2$-value), the predictive value of this test for the accuracy of the estimate derived from the meta-analysis is unknown. Furthermore, there is no uniform approach to dealing with heterogeneity. Multiple strategies have been proposed [3], and there are many examples of meta-analyses being performed in the presence of substantial heterogeneity. In this study, we investigated, by using a simulation model, the accuracy of the $I^2$-statistic in the assessment and quantification of heterogeneity, and how heterogeneity across studies relates to the predictive value of meta-analyses. First, we briefly explain the concepts of heterogeneity. Subsequently, the objectives and the results of our simulation model are presented. Finally, we illustrate and clarify our findings by presenting common scenarios including several examples of meta-analyses evaluating different interventions published in the field of infectious diseases and critical-care medicine.

### Heterogeneity

Heterogeneity across studies includes all differences between individual studies related to, among other factors, study design, populations included, treatment strategies, and outcomes. For simplicity, we distinguish two types of heterogeneity: ‘owing to chance’ and ‘systematic’.

Even when the strictest selection criteria for study inclusion are used, it is impossible to avoid some kind of heterogeneity between studies performed under different conditions. In fact, even in the hypothetical situation of a single study being executed multiple times under exactly the same conditions, the outcome will, owing to chance events, not be exactly the same for each evaluation. In addition to this unavoidable heterogeneity owing to chance, there is a possibility of heterogeneity owing to systematic differences between the studies, such as differences in study design, patient populations, diagnostic methods, application of interventions, or definitions of outcome. Some level of heterogeneity can be avoided by using strict criteria of study selection, based on design (i.e. only double-blind randomized trials instead of any randomized trial), populations (only mechanically ventilated trauma patients instead of all types of mechanically ventilated patients), and outcomes (i.e. only day 28 mortality instead of mortality measured at different time-points). Therefore, although heterogeneity can be avoided to some extent, it can never be prevented completely. However, the predictive value of meta-analyses is unknown in the case of systematic heterogeneity.

Several methods have been proposed for quantification of heterogeneity in meta-analyses [3]. Such a test examines the null hypothesis that all studies have evaluated the same effect. Cochran’s Q reflects the sum of the squared deviations of the study’s estimate from the overall pooled estimate, weighing each study’s contribution in the same way. However, this test is poor in detecting true heterogeneity, especially when small numbers of studies are being dealt with.

$I^2$ reflects the percentage of total variation across studies that is attributable to heterogeneity rather than chance, and is calculated from Cochran’s Q as $100 \times (Q – degrees\ of\ freedom)/Q$. Negative $I^2$-values are considered as 0%, which indicates no observed heterogeneity. Heterogeneity can be quantified as low, moderate, and high, with upper limits of 25%, 50% and 75% for $I^2$, respectively. Calculation of $I^2$ has now become the standard way of reporting heterogeneity in all Cochrane reviews [2,3]. Interestingly, $I^2$ is almost always reported as a single value without a 95% CI, although these areas can be wide, demonstrating the inherent uncertainty of this value [4]. It is neither possible to quantify the exact level of...
heterogeneity across studies nor to distinguish the contribution of chance and systematic heterogeneity.

Adequately dealing with heterogeneity is often difficult, although some guidelines are provided [3]. Important questions are whether heterogeneity is too large for a meaningful meta-analysis, and what particular model should be used for calculation of the pooled estimate. For the latter, two models are used [3]: (i) the fixed-effects model, which assumes that all of the included studies are estimating the true effect, and that variation in findings among the studies is therefore attributable to chance only; and (ii) a random-effects model, which assumes that the effects estimated in the different studies follow a distribution, resulting in a wider CI of the pooled estimate.

Methods

In our simulation model meta-analyses were performed of simulated randomized controlled trials evaluating a certain intervention that had varying amounts of heterogeneity, with the following assumptions. We assume an intervention with 25% efficacy (e.g. mortality reduction), and that the outcome occurs in 15% of the population when the intervention is performed (and thus in 20% without intervention). The amount of systematic heterogeneity is expressed by $\sigma$, which is 0 in the absence of systematic heterogeneity. Monte Carlo simulations are used to perform multiple meta-analyses, each including ten studies with two groups of 100 patients. For each simulated study, the expected mortalities are 0.20 and 0.15 in the control and intervention arms, respectively. In the intervention group, however, we deliberately modify the amount of heterogeneity (see methods in Supporting Information). Meta-analysis results were compared with the true average relative risk of simulated studies and with a (simulated) reference study with an infinite number of patients, which therefore always yields the average incidence of 0.15. We determined: (i) whether the 95% CI of the meta-analysis contains the true relative risk (0.75) in at least 81% of the simulations, and with low systematic heterogeneity between intervention arms ($\sigma < 0.05$), the 95% CI contains the true relative risk in at least 93% of all experiments (Fig. 1a). In fact, the CI obtained from a random-effects meta-analysis has the desired properties as long as the differences between studies are more or less normally distributed (see methods in Supporting Information).

However, the 95% CI of a meta-analysis does not accurately predict the outcome of a single large study. In a series of figures (Fig. 1 and Figs S2–S4), we have depicted the associations

Results

With a random-effects meta-analysis, the 95% CI of the effect estimate contains the true relative risk (0.75) in at least 81% of the simulations, and with low systematic heterogeneity between intervention arms ($\sigma < 0.05$), the 95% CI contains the true relative risk in at least 93% of all experiments (Fig. 1a). In fact, the CI obtained from a random-effects meta-analysis has the desired properties as long as the differences between studies are more or less normally distributed (see methods in Supporting Information).

However, the 95% CI of a meta-analysis does not accurately predict the outcome of a single large study. In a series of figures (Fig. 1 and Figs S2–S4), we have depicted the associations...
between the accuracy of estimates derived from meta-analyses and the chance that the result of the 11th study falls within the 95% CI of the meta-analysis estimate, as well as the chance that it contains the true relative risk, with increasing amounts of heterogeneity. As expected, \( I^2 \) increases and the likelihood of drawing correct inferences from a meta-analysis decreases with increasing heterogeneity (Fig. 1). With a random-effects model, the chances of correct estimates are higher than with a fixed-effects model, especially when heterogeneity increases (Fig. 1).

Surprisingly, however, in the case of low levels of heterogeneity (\( \sigma \) close to 0), \( I^2 \) appears to be not predictive of the accuracy of the meta-analysis result. With low amounts of heterogeneity (\( \sigma \) close to 0%), even fixed-effects meta-analyses with \( I^2 > 75\% \) yield highly accurate results (Fig. S2). With high levels of heterogeneity, even meta-analyses with low \( I^2 \)-values have low predictive values (Fig. S2). With a random-effects model, the width of the 95% CIs increases with increasing \( I^2 \), which increases the likelihood of the 11th study result falling within the CI limits (Fig. S3). However, the likelihood of obtaining low \( I^2 \)-values also depends on the amount of heterogeneity (Fig. S4). In the absence of any heterogeneity (a pure theoretical option), the chance of finding a high \( I^2 \) (>50%) is very low, but this rapidly increases with increasing levels of heterogeneity. However, even with high heterogeneity (large \( \sigma \)), low \( I^2 \)-values can be derived. All simulations were repeated with an inverse variance method (instead of the Mantel–Haenszel method), leading to identical results (data not shown).

How do these results relate to the daily practice of performing meta-analyses? We propose six different scenarios, depending on the amount of expected methodological and clinical heterogeneity (low, high, and unknown) and the \( I^2 \)-values (high and low) that may occur when a meta-analysis is performed (Fig. 2).

**Scenarios 1 and 2**
These scenarios relate to situations in which considerable heterogeneity is probably attributable to differences between studies (e.g., differences in study design, patient populations, diagnostic methods, application of interventions, or definitions of outcome). Calculated \( I^2 \)-values can be low (scenario 1) or high (scenario 2), but even low \( I^2 \)-values (scenario 1) can be associated with a low predictive values of meta-analysis results (Fig. S2). A high \( I^2 \) (scenario 2) is intuitively correct, considering the obvious amount of clinical and methodological heterogeneity, and only confirms what was already expected. The high \( I^2 \), however, will have broad CIs if a random-effects method is used, which will contain the true average relative risk in approximately 95% of the simulations. However, the 95% CI of the meta-analysis will not contain the result of the large 11th study in at least 95% of the simulations. Thus, calculation of \( I^2 \) in both scenarios is of limited value: in scenario 1, a low \( I^2 \) can be obtained, which is associated with low predictive values of meta-analysis results; in scenario 2, it only confirms what was already known (i.e. high levels of heterogeneity with low accuracy of the meta-analysis).

**Scenarios 3 and 4**
These scenarios relate to situations where, based on the similarity of the included studies, a low amount of heterogeneity is expected. The chance of obtaining a high \( I^2 \) (scenario 4) is low (Fig. S4), but the accuracy of the meta-analysis result will still be as high as with low \( I^2 \)-values (scenario 3) (Fig. S2). Therefore, with extremely low levels of heterogeneity, \( I^2 \)-values are not informative regarding the accuracy of meta-analyses.

**Scenarios 5 and 6**
These scenarios relate to (probably frequent) situations in which differences between studies exist, but where the impact of these differences on the pooled estimates are unknown. In such situations, a reliable statistical method to quantify the amount of heterogeneity is especially needed. In our simulation studies, high \( I^2 \)-values (>75%) (scenario 6) are predictive for the presence of heterogeneity (Fig. S4) and low predictive values of the estimates derived (Fig. S2). However, low \( I^2 \)-values (scenario 5) correspond to a wide range of systematic heterogeneity levels, and thus to a high level of uncertainty about the predictive value of meta-analysis results. Even when \( I^2 = 0 \), systematic heterogeneity can exist, which will reduce the reliability of the pooled estimate. In real life, levels of heterogeneity of most meta-analyses will be unknown. In these scenarios, \( I^2 \) determination may help to identify estimates with low predictive values (high \( I^2 \)). In this situation, the results of a meta-analysis will be unreliable. With low \( I^2 \)-values and

**FIG. 2.** Clinical scenarios.
unknown levels of clinical and methodological heterogeneity, predictive values of pooled estimates may range extensively, and findings should be interpreted with caution.

We illustrate only the first two scenarios, as scenarios 3 and 4 are scarce and scenarios 5 and 6 unknown, with two clinical examples of meta-analyses published in the field of infectious diseases and critical-care medicine.

Clinical example I
The effectiveness of selective decontamination of the digestive tract (SDD) has been evaluated in several meta-analyses (Table 1) [5–11]. These meta-analyses included different studies, partly because not all studies were available at the time of preparation, or because different selection criteria were applied. There were also differences in the aggregate data used per study; some meta-analyses preferably used intention-to-treat data, whereas others preferentially used the data of patients with a length of stay of at least 48 h. Also, some used the hospital mortality data when available, whereas others only used intensive-care unit (ICU) mortality. Despite these differences, the pooled estimate of efficacy of SDD in reducing mortality remained more or less stable with ORs around 0.80, being statistically significant in the most recent analyses. Silvestri et al. [10] summarized the characteristics of the 30 studies included in their meta-analysis. Mortality in control patients ranged from 3% to 58% (average of 25% with a standard deviation of 15%), the methodological study quality ranged (on a scale of 16) from 6 to 14 (average of 9 ± 5), 11 different patient populations were studied, 11 different intravenous medications were tested (including no prophylaxis), 11 of 30 studies used intravenous prophylaxis in control patients, two studies evaluated oropharyngeal decontamination only, and three evaluated intestinal decontamination only. Nevertheless, some used the hospital mortality data when available, whereas others only used intensive-care unit (ICU) mortality. Despite these differences, the calculated \( I^2 \) in their meta-analysis is 0%, and similarly as low as for the other meta-analyses. On the basis of study characteristics, heterogeneity was expected to be high, but this was not reflected by the obtained \( I^2 \)-values, resembling scenario 1. After publication of these meta-analyses, the effects of SDD on patient outcome were reported from a multicentre trial including more patients \( (n = 5939) \) than in all of the studies included in the most recent meta-analysis [12]. In this multicentre study, SDD was, as compared with standard care, associated with a 13% reduction in day 28 mortality, which corresponded to an adjusted OR of 0.83. This result was remarkably similar to the results obtained in previous meta-analyses. Thus, despite obvious methodological and clinical differences between individual studies, the meta-analyses seemed to have accurately predicted the effects of SDD.

Clinical example II
In another meta-analysis, the effects of weaning protocols on the duration of mechanical ventilation in critically ill adult patients was determined [13]. Eleven studies, both randomized and quasi-randomized controlled trials, were selected, evaluating 1971 patients admitted to seven different types of ICU. Only two studies used the same weaning protocol, and the ‘usual care’ in the control group comprised a wide variety of practices. The authors used fixed-effects models for meta-analysis, and a random-effects model in the case of statistical heterogeneity (defined as \( I^2 > 50\% \) and/or chi-square statistic \( p < 0.05 \)). The primary outcome was the duration of mechanical ventilation with and without the weaning protocol, which was estimated (with random-effects model) as mean log \( 0.29 \) (95% CI \( -0.5 \) to \( -0.09 \)). However, a substantial amount of heterogeneity was quantified with \( I^2 (I^2 = 76\%) \). Subgroup analyses to assess the impact of type of ICU were small (two to four studies), and did not reduce heterogeneity as indicated with the statistical test. Several secondary outcomes were tested; no heterogeneity was indicated (\( I^2 \) was estimated as low) in the analyses concerning hospital mortality (pooled estimate of 1.10, 95% CI 0.86–1.41, \( I^2 = 0\% \), \( p = 0.46 \)) and length of stay in the ICU (pooled estimate of \( -0.11 \), 95% CI \( -0.21 \) to \( -0.02 \), \( I^2 = 0\% \), \( p = 0.45 \)), and marked heterogeneity was indicated in the analyses of ICU mortality (pooled estimate of 0.98, 95% CI 0.48–2.02, \( I^2 = 57\% \), \( p = 0.07 \)) and duration of weaning (pooled estimate of \( -1.52 \), 95% CI \( -2.66 \) to \( -0.37 \), \( I^2 = 97\% \), \( p < 0.001 \)). The authors concluded the following: ‘compared with usual care, use of weaning protocols can reduce the duration of mechanical ventilation by 25%, weaning duration by 78% and length of stay in the ICU by 10%. As there was significant heterogeneity in included trials and most were conducted in the US, these findings might not be generalisable.’ Indeed, heterogeneity was

### Table 1. Meta-analyses published that have evaluated selective digestive decontamination

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>No. of studies</th>
<th>OR (95% CI)</th>
<th>( I^2 ) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vandenbroucke-Grauls</td>
<td>1991</td>
<td>7</td>
<td>0.70 (0.45–1.09)</td>
<td>0</td>
</tr>
<tr>
<td>SDD CTG</td>
<td>1993</td>
<td>15</td>
<td>0.80 (0.67–0.97)</td>
<td>0</td>
</tr>
<tr>
<td>Heyland</td>
<td>1994</td>
<td>24</td>
<td>0.83 (0.71–0.98)</td>
<td>0</td>
</tr>
<tr>
<td>Kallaf</td>
<td>1994</td>
<td>16</td>
<td>0.88 (0.72–1.08)</td>
<td>0</td>
</tr>
<tr>
<td>Hurley</td>
<td>1995</td>
<td>26</td>
<td>0.86 (0.74–0.99)</td>
<td>5</td>
</tr>
<tr>
<td>D’Amico</td>
<td>1998</td>
<td>17</td>
<td>0.80 (0.69–0.93)</td>
<td>10</td>
</tr>
<tr>
<td>Silvestri</td>
<td>2007</td>
<td>30</td>
<td>0.80 (0.69–0.94)</td>
<td>0</td>
</tr>
</tbody>
</table>

Pooled ORs as provided in the meta-analysis or calculated from the information in the manuscript of the meta-analyses. \( I^2 \), when unavailable, was calculated by use of the chi-squared statistic and degrees of freedom. The meta-analysis of SDD CTG was updated; here, only the results of the first publication are provided.
expected to be high in this meta-analyses (owing to differences in intervention, control groups, and patient population), and this was confirmed by a high $I^2$ in many analyses, thereby resembling scenario 2. The estimates obtained should therefore be interpreted with extreme caution.

**Discussion**

In this study, we have demonstrated the crucial importance of study selection (or, in other words, of minimizing clinical and methodological heterogeneity) for the accuracy of pooled estimates derived from meta-analyses. Quantifying statistical heterogeneity through $I^2$-statistics can be helpful in some scenarios (when the amount of heterogeneity is unknown and $I^2$ is high), but is of no help in other scenarios (at the extremes of heterogeneity levels and when the amount of heterogeneity is unknown and $I^2$ is low). Our findings underscore the need for a critical appraisal of meta-analyses before their results are accepted, and underscore the huge responsibility of meta-analysts (and peer reviewers and editors) in adequately performing, interpreting and reporting of meta-analyses.

The reliability of $I^2$-statistics in quantifying levels of heterogeneity has been questioned before, albeit without determination of its association with estimate accuracy. Huedo-Medina et al. [14] demonstrated that $I^2$ suffers from low statistical power, potentially leading to misleading results, when the number of studies is small. Ioannidis et al. [4] emphasized that, like any metric, $I^2$ has some uncertainty that can be expressed in 95% CIs. In Cochrane meta-analyses with $I^2$-values of ≤25%, 83% of these values had upper 95% CIs that crossed into the range of large heterogeneity (≥50%). Even when $I^2$ was 0%, 81% had CIs exceeding 50%. However, these intervals are still rarely provided.

As meta-analyses have become so important in evidence-based medicine, their results should be reliable and accurate. Our findings demonstrate that heterogeneity importantly influences both aspects. As yet, there are no reliable methods with which to quantify the amount of clinical and methodological heterogeneity, and careful selection of appropriate studies is the only tool for deriving correct inferences from meta-analyses. Unfortunately, this selection will always be, at least to some extent, subjective. Our findings demonstrate that determination of $I^2$ is of little value at the extremes of heterogeneity, and it would be helpful to derive criteria for categorizing meta-analyses into either low or high levels of clinical and methodological heterogeneity. The consequence would be that meta-analyses with high levels of clinical and methodological heterogeneity should not provide a pooled estimate (as the level of accuracy will always be low, regardless of $I^2$) and that meta-analyses with low levels of heterogeneity should not provide an estimate of $I^2$. In real life, however, the levels of heterogeneity of most meta-analyses will be unknown. In such scenarios, $I^2$ determination may help to identify estimates with low predictive values (high $I^2$), for which we recommend that pooled estimates should not be provided. With low $I^2$-values and unknown levels of clinical and methodological heterogeneity, predictive values of pooled estimates may range extensively, and findings should be interpreted with caution. Objective methods to quantify the levels of clinical and methodological heterogeneity are urgently needed to allow reliable determination of the accuracy of meta-analyses. Until that time, we propose that investigators describe the pretest likelihood of clinical and methodological heterogeneity and carefully discuss the potential effects on study results.

**Transparency Declaration**

The authors declare no conflict of interest.

**Supporting Information**

Additional Supporting Information may be found in the online version of this article:

- **Figure S1.** Beta distribution.
- **Figure S2.** Results of Monte Carlo simulation study: fixed-effect analysis.
- **Figure S3.** Results of Monte Carlo simulation study: random-effects analysis.
- **Figure S4.** Results of Monte Carlo simulation study: chance of observing a certain level of the $I^2$-statistic.

**References**


