REVIEW 10.1111/1469-0691.12494

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

W. G. Melsen[1], M. C. J. Bootsma[1,2], M. M. Rovers[3] and M. J. M. Bonten[1,4]

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

Corresponding authors: W. G. Melsen, Heidelberglaan 100, Utrecht, 3584 CX, the Netherlands
E-mail: W. G.Melsen@umcutrecht.nl
and
M. J. M. Bonten, Heidelberglaan 100, 3584 CX Utrecht, the Netherlands
E-mail: m.j.m.bonten@umcutrecht.nl

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 $100\% \times (Q - \text{degrees of freedom})/Q$. 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