PharmacoEconomics (2013) 31:719–730
DOI 10.1007/s40273-013-0067-0

ORIGINAL RESEARCH ARTICLE

A Choice That Matters?

Simulation Study on the Impact of Direct Meta-Analysis Methods on Health Economic Outcomes

Pepijn Vemer · Maiwenn J. Al · Mark Oppe · Maureen P. M. H. Rutten-van Mölken

Published online: 5 June 2013

Abstract

Background Decision-analytic cost-effectiveness (CE) models combine many different parameters like transition probabilities, event probabilities, utilities and costs, which are often obtained after meta-analysis. The method of meta-analysis may affect the CE estimate.

Aim Our aim was to perform a simulation study that compares the performance of different methods of meta-analysis, especially with respect to model-based health economic (HE) outcomes.

Methods A reference patient population of 50,000 was simulated from which sets of samples were drawn. Each sample drawn represented a clinical trial comparing two fictitious interventions. In several scenarios, the heterogeneity between these trials was varied, by drawing one or more of the trials from predefined subpopulations. Parameter estimates from these trials were combined using frequentist fixed (FFE) and random effects (FRE), and Bayesian fixed (BFE) and random effects (BRE) meta-analysis. The pooled parameter estimates were entered into a probabilistic cost-effectiveness Markov model. The four methods of meta-analysis resulted in different parameter estimates and HE outcomes, which were compared with the true values in the reference population. Performance statistics were: (1) the percentage of repetitions that the confidence interval of the probabilistic sensitivity analysis covers the true value (coverage), (2) the difference between the estimated and true value (bias), (3) the mean absolute value of the bias (MAD) and (4) the percentage of repetitions that result in a statistically significant difference between the two interventions (statistical power). As the differences between methods could be due to chance, we repeated every step of the analysis 1,000 times to study whether differences were systematic.

Results FFE, FRE and BFE lead to different parameter estimates, but, when entered into the model, they do not lead to large differences in the point estimates of the HE outcomes, even in scenarios where we built in heterogeneity. Random effects methods do not necessarily reduce bias when heterogeneity is added to the trials, and may even increase bias in certain situations. BRE tends to overestimate uncertainty reflected in the CE acceptability curve.

Conclusion FFE, FRE and BFE lead to comparable HE outcomes. BRE tends to overestimate uncertainty. Based on this study, we recommend FRE as the preferred method of meta-analysis.

Key Points for Decision Makers

1. To aid decision making, available evidence is often structured in a probabilistic decision-analytic model. Meta-analysis combines all available evidence in a wide range of model parameters.
2. A Bayesian random effects approach of meta-analysis is not recommended when only a few sources of evidence are available, as it may overestimate uncertainty and yield a larger probability that a new treatment is rejected or that more research is asked for, where it might not be necessary.
3. With only a few differences between the other three methods we compared, we recommend a frequentist random effects approach as the preferred method of meta-analysis.

Electronic supplementary material The online version of this article (doi:10.1007/s40273-013-0067-0) contains supplementary material, which is available to authorized users.

P. Vemer (✉) · M. J. Al · M. Oppe ·
M. P. M. H. Rutten-van Mölken
Institute for Medical Technology Assessment (iMTA), Erasmus
University, PO Box 1738, 3000 DR Rotterdam, The Netherlands
e-mail: pepijnvemer@gmail.com

1 Introduction

In 2006, the Netherlands implemented a policy of conditional, temporary reimbursement of potentially innovative, but expensive hospital drugs [1]. Additional hospital funding is provided on the condition that outcomes research is performed to show further evidence of the value of the new drugs. The final reimbursement decision is made based on all evidence available, after 4 years. A systematic approach to aid decision making is called comprehensive decision modelling, in which available evidence is structured in a probabilistic decision-analytic model [2, 3]. Meta-analysis is one step in this process, and is used to combine all available evidence in model parameters. A wide range of model parameters need to be estimated, from transition probabilities to costs and utility values [4].

Many different methods of meta-analysis exist, and many authors have compared them (e.g. [5–8]). They have shown that the choice of method can considerably affect parameter estimates. These comparisons concentrated on the impact of the method of meta-analysis on the estimate of a single treatment effect, for example a risk ratio (RR). However, in the probabilistic models used in economic evaluations we need to estimate many different parameters, including the baseline value of each model parameter in the comparator group. Altogether, the method of meta-analysis to obtain these parameters may considerably affect the final cost-effectiveness (CE) estimates.

Our group has previously investigated the effect of four different methods of meta-analysis on model-based CE estimates [9]. Although we found considerable differences, there was no way of knowing which of the methods was best, because we had no ‘truth’ to which we could compare our results. That is, we only had data from different samples of the total patient population, not of the population itself. To overcome this problem we performed a simulation study, in which we created a reference population, which reflected the value that should be obtained by the different methods. We then proceeded by drawing sets of samples from this population, mimicking sets of clinical trials, and combined these trial estimates. Each method of meta-analysis generated a separate set of pooled parameters. We filled a health economic (HE) model with these different sets of parameters and investigated whether there were systematic differences between the meta-analysis methods by comparing the outcomes of the sets of samples with the outcomes of the reference population. We were especially interested in the impact on the differences in costs and quality-adjusted life years (QALYs), the incremental CE ratio and the CE acceptability curve.

The available methods of meta-analysis can be divided into two groups, namely direct and indirect methods. Direct methods of meta-analysis combine evidence from trials that compare the two interventions of interest directly. In the absence of head-to-head studies, or with the availability of both direct and indirect evidence, indirect methods of meta-analysis come into play. Methods of indirect meta-analysis are compared in a separate manuscript (under preparation) by the same authors. We therefore focus here on direct meta-analysis methods.

2 Methods

2.1 Simulation Study

The simulation comprised several steps, shown in Fig. 1. In step 1 (Create reference population), we simulated a reference patient population ($n = 50{,}000$), including individual patient-level disease progression using one of two fictitious treatments. The mean values of the parameters and HE outcomes, as calculated from the entire population, are reference values to which we compared the estimates of the meta-analyses. In other words, they represented the ‘truth’ and are referred to as *reference parameters* and *reference outcomes*. Parameters included transition probabilities, probabilities to experience an event, maintenance costs, utilities and costs and utility-decrements due to an event. HE outcomes included the total number of QALYs, life years (LY) and events, intervention and maintenance costs, and the incremental CE ratio (ICER).

In step 2 (Trial selection), we sampled trials from the reference population, comparing the two treatments. For each of the trials we calculated the parameters that are needed as input for the HE model, called *trial parameters*. In step 3 (Meta-analysis), we pooled the trial parameters using several methods of meta-analysis. These methods are explained in detail in Sect. 2.3. The combined estimates are called *model parameters*. For each model parameter, both mean and appropriate dispersion measures were calculated. We used a disease progression model in step 4 (CE modeling), filled first with a set of model parameters obtained by the first method of meta-analysis. A probabilistic sensitivity analysis (PSA; 1,100 iterations) was run and the HE outcomes, called *model outcomes*, were collected. This process was repeated with model parameters obtained from each of the methods of meta-analysis.

Differences in model outcomes could be due to chance, i.e. the particular set of trials that was drawn. In order to study whether there was a systematic difference between the methods of meta-analysis, we repeated steps 2 to 4 in step 5 (Repeat), further referred to as 1,000 *repetitions*.

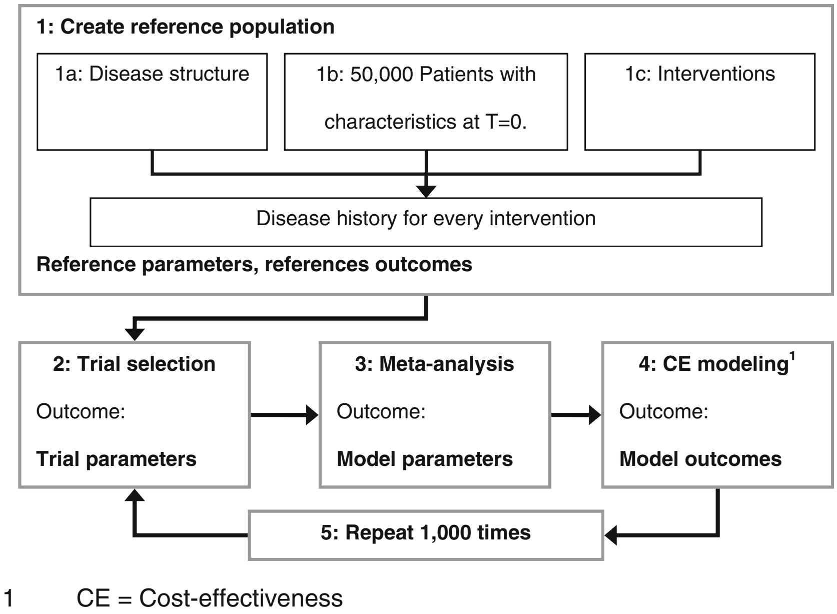
1: Create reference population
1a: Disease structure
1b: 50,000 Patients with characteristics at T=0.
1c: Interventions
Disease history for every intervention
Reference parameters, references outcomes
2: Trial selection
Outcome:
Trial parameters
3: Meta-analysis
Outcome:
Model parameters
4: CE modeling¹
Outcome:
Model outcomes
5: Repeat 1,000 times
1 CE = Cost-effectiveness

Fig. 1 Design of the simulation study

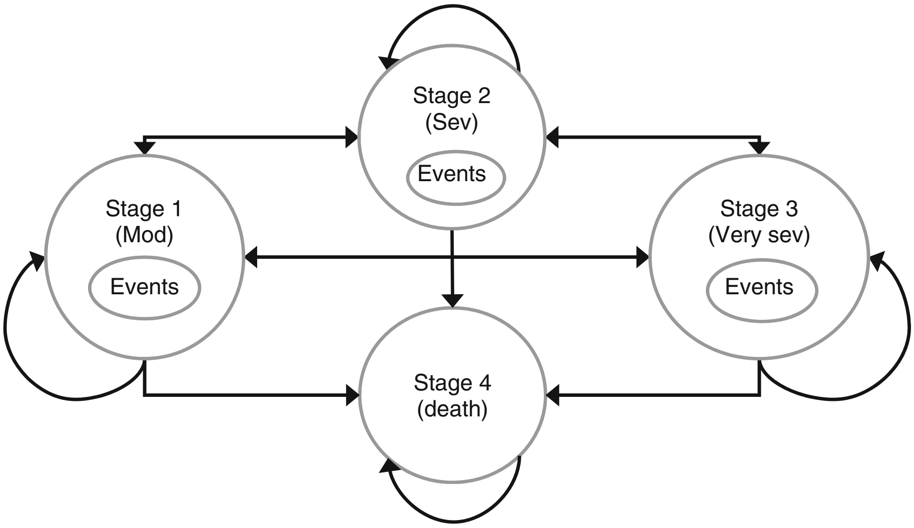

Fig. 2 Markov model of the chronic disease

2.2 Disease and Model Structure

The modelled disease was a progressive, chronic disease (Fig. 2), with events during which symptoms worsen considerably. The disease was simulated using a Markov model with four stages: moderate, severe, and very severe disease, and death.

For each patient in the reference population, we simulated their disease progression. We did this by first defining the reference disease progression (RDP), which can be thought of as the disease progression of an untreated, base-case patient. It consists of a set of distributions for each reference parameter (Table 1). Next, these distributions were modified based on individual patient characteristics—gender, age, developed/developing country, body-mass index (BMI) and smoking status and interventions. These characteristics made it possible for us to add heterogeneity to trials in relevant scenarios, by sampling from sub-populations. In this manner we have simulated a heterogeneous population of individual patients.

How patient characteristics and interventions influenced the RDP is stated in an online appendix (see electronic supplementary material). In short, male patients have a higher probability to move to a worse disease stage than female patients. Older patients have a higher probability to move to a worse disease stage than younger patients; they have higher costs and a wider spread in quality-of-life weights. Patients from developing countries have lower maintenance costs than patients from developed countries. Patients with a higher BMI have a higher probability to move to a worse disease stage than patients with lower BMI; they also have a higher probability of an event,

Table 1 Characteristics of the simulated patient population

Size simulated cohort	50,000
Starting disease stage	5/8 in moderate, 2/8 in severe and 1/8 in very severe
Gender	50 % male, 50 % female
Age in years	18–34; 35–64; 65+
	Determined by a random draw from a uniform distribution from 18 to 75
Developed/ developing country	50 % from developed countries, 50 % developing countries
Body Mass Index (BMI)	<25 kg/m^2 (average or low); 25–30 kg/m^2 (high); >30 kg/m^2 (obese)
	Determined by a random draw from a normal distribution with mean 23 and standard deviation of 4
Smoking status	30 % smokers, 70 % non-smokers

higher maintenance costs and lower quality of life. Patients who smoke have a higher probability to move to a worse disease stage and a higher probability of an event, than patients who do not smoke.

Interventions influence the RDP in the same manner as patient characteristics do. For each patient in the reference population, we simulated their disease progression twice: once receiving Usual Care and once receiving the New Intervention. Usual Care is a drug that decreases the probability of disease progression compared with the RDP, at €60 per monthly cycle. New Intervention, the focus of the HE analysis, decreases the probability of disease progression, more so than Usual Care, plus it increases the probability of moving to a better disease stage and decreases the probability of an event. The costs were set at €350 per monthly cycle. In the HE model, probabilities for the New Intervention are modeled as a RR, with the estimated probabilities for the Usual Care as a baseline.

Changes to reference parameters were additive across patient characteristics and interventions. For example, a female patient aged 35–64 years who used the New Intervention had a monthly probability to die in the very severe disease stage of 10 % (the probability within the RDP) – 2 % (modification for gender) + 4 % (modification for age) − 3 % (modification for New Intervention) = 9 %.

Table 2 shows the reference outcomes when applying the two interventions to the entire patient population. They represent the ‘truth’ with which the outcomes of the meta-analyses were compared.

The structure of the HE model mirrors the disease progression in the reference population; in other words, there was no structural uncertainty. The time horizon of the HE model was 1 year and the cycle length 1 month. We assumed that data in the trials were collected each month during 1 year. We have not applied discounting. Simulation and modelling were performed using SAS 9.2 and WinBUGS 1.4.3.

2.3 Scenarios

The number and size of the trials sampled in step 2: Trial selection was varied in scenarios, as well as the amount of heterogeneity between trials. Heterogeneity in the meta-analysis literature is any kind of variability between different studies [10]. Trial heterogeneity is different from patient heterogeneity, which is the difference between patients that can be adequately explained by patient characteristics. Table 3 shows the different scenarios that were investigated. The last column of Table 3 described the impact of the non-randomly drawn trials on the trial parameters. We will focus mainly on the three scenarios in shaded rows, namely 1, 4 and 7. The other scenarios will be discussed in the discussion section of this paper.

2.4 Methods of Meta-Analysis

In our study, we compared four widely used methods of meta-analysis: frequentist fixed effects (FFE), frequentist random effects (FRE), Bayesian fixed effects (BFE) and Bayesian random effects (BRE). The FFE and FRE were based on the Inverse Variance method, which can be used for meta-analysis of both continuous and dichotomous data [11]. The pooled effect estimate for the FFE is calculated as a weighted average of the individual study estimates, using the inverse of the squared standard error (SE) of the effect estimates as weights. Thus, studies with a smaller SE, typically larger studies, are given more weight than studies with a larger SE. For the FRE, we used the DerSimonian–Laird method [11]. It was developed for situations where there is heterogeneity between study results, caused, for example, by differences in patient population or study design. It incorporates an estimate of the between-study heterogeneity into the weights. It is assumed that all studies are samples drawn from a pool of all possible studies, i.e. the population [10]. The goal is to estimate the mean of this population. The true parameter value may be study specific and can vary across studies.

Both the FFE and FRE assume that the weights are known. With little or no heterogeneity among the studies, the FFE and FRE will give identical results [10]. With heterogeneity present, confidence intervals will be wider for the FRE and claims of statistical significance will be more conservative. The point estimate of the parameter might also be different. We report the I^2-statistic as a measure of

Table 2 Reference outcomes, per patient per 12 cycles/months − Mean (standard deviation)

Variables	Usual care	New intervention	Difference
QALYs	0.485 (0.232)	0.540 (0.231)	0.054
LYs	0.740 (0.328)	0.786 (0.313)	0.046
Intervention costs (€)	533 (236.24)	3,300 (1,310)	2,770
Maintenance costs (€)	3,260 (2,080)	3,070 (1,810)	−180
Event costs (€)	2,330 (2,610)	1,260 (1,780)	−1,070
Total costs (€)	6,120 (4,340)	7,630 (3,830)	1,520
Number of cycles in:			
Moderate disease	5.171 (3.750)	6.209 (3.965)	1.038
Severe disease	2.477 (2.512)	2.313 (2.507)	−0.164
Very severe disease	1.238 (1.850)	0.911 (1.554)	−0.327
Death	3.114 (3.937)	2.567 (3.751)	−0.547
Number of events	1.160 (1.259)	0.630 (0.856)	−0.530
Proportion surviving (%)	49.9	58.3	8.4
ICER, total costs per QALY (€)			28,020

LY life year, *QALY* quality-adjusted LY, *ICER* incremental cost-effectiveness ratio

Table 3 Overview of different scenarios in the simulation study[a]

Scenario	Number of trials	Number of patients per intervention arm	Added heterogeneity with effect on disease progression
1	5	All trials 500	-
2	5	Trial 1 and 2: 500, trial 3: 100, trial 4: 250, trial 5: 1,000	-
3	10	All trials 250	-
4	5	All trials 500	Trial 5 has relatively old patients, more smokers and more obese patients, which leads to more rapid disease deterioration, higher probability of events, higher maintenance costs, lower quality of life.
5	5	All trials 500	Trial 2 has relatively young patients, which leads to slower disease deterioration Trial 4 has only patients from developing countries, which leads to lower maintenance costs Trial 5: the same as in scenario 4
6	5	Trials have different sample sizes, the same as in scenario 2	The same as in scenario 5
7	5	Trials have different sample sizes, the same as in scenario 2	Trials 2, 4 and 5 have relatively old patients, more smokers and more obese patients, which leads to more rapid disease deterioration, higher probability of events, higher maintenance costs, lower quality of life.

[a] The four scenarios that are not in shaded rows are only discussed in the discussion section

heterogeneity [12], which can be interpreted as the proportion of the total variation in the pooled estimates that is due to heterogeneity between studies. When the amount of between-trial heterogeneity increases compared with the within-trial variance, then the I^2 also increases. Higgins and Green [8] provide a rough guide to the interpretation of I^2. Above 30 % "may represent moderate heterogeneity"; above 50 % "may represent substantial heterogeneity".

The BFE method requires the data from the different trials, the definition of a prior for the parameter to be synthesized and a likelihood linking both [9, 13]. We used a binomial likelihood function to model the total number of transitions, with a flat beta prior; and a normal likelihood function for all other parameters, with a flat normal prior centered on 0 and a precision of 1.0E−6. When specifying the BRE method, prior distributions need also be defined for the between-trial heterogeneity [9, 13, 14]. We used the inverse of a squared uniform distribution from 0 to 10. Other likelihoods and priors were as in the BFE. Before simulation started, we tested several priors and could find no meaningful differences.

Conceptually, confidence intervals in frequentist statistics and credibility intervals in Bayesian statistics have very different interpretations (see for example [15, 16]). However, for convenience and legibility, we abbreviate both as CI. For each pooled parameter estimate, we report the mean and the 95 % CI.

We performed meta-analysis on all baseline values (transition probabilities, utilities, etc.) using information from the Usual Care intervention arms. In addition, we performed meta-analysis on all effect measures (RR), using data on the difference between the New Intervention and Usual Care. Interested readers may request code on both the simulation study and the methods of meta-analysis from the corresponding author.

2.5 Comparing Performance

When judging the performance of the methods of meta-analysis, we assumed that a researcher doing a meta-analysis aims to estimate the CE of the New Intervention compared with Usual Care in the entire patient population, not a specific subgroup. We further assumed that a researcher is unaware of the fact that heterogeneity, when present, was caused by sampling from subgroups (i.e. they do not know we deliberately built in heterogeneity). To the researcher, the heterogeneity might either be caused by random sampling or unobserved differences between the trials in terms of patient characteristics, setting or other elements that could affect the parameter estimates. These assumptions are made because, if these differences in design are known to the researcher, either the trials would not be synthesized at all, or a way has to be found to control for these differences. Hence, these assumptions made it possible to judge the performance of the different methods of meta-analysis by comparing model parameters and outcomes with the reference values.

The statistical performance of the different methods was judged by calculating the coverage, bias, mean absolute deviation (MAD) and statistical power. Coverage is the percentage of all repetitions that the simulated CI covered the 'truth'. Since the coverage is based on 95 % CIs, we expect that, if all trials are drawn randomly, the coverage should on average be close to 95 % [5]. The observed coverage was compared to this benchmark. Assuming that we have an unbiased point estimate, if the coverage is below 95 %, the model does not take into account all uncertainty. If the coverage is above 95 %, it has accounted for too much uncertainty. In this study, if the coverage was smaller than 90 %, we say the method underestimated uncertainty; if the coverage was higher than 98 % the method overestimated uncertainty. Bias is expressed as the difference between the point estimate in the simulated data set and the true population value, averaged over all repetitions. The MAD is the average, over all repetitions, of the absolute value of the bias. The MAD indicates how far the estimated value was from the 'truth', regardless of whether it was too high or too low. For HE outcomes, we also calculated statistical power, expressed as the percentage of all repetitions where the simulated result yields a statistically significant difference between treatments.

3 Results

3.1 Model Parameters for One Set of Trials

Figure 3 compares the methods for scenarios 1, 4 and 7, using only the first repetition. From bottom to top, we compare the different meta-analysis models for the seven scenarios. Each dot represents the point estimate for the model parameter, in this case the transition probability from severe to very severe disease, and the bars the estimated CIs. At the bottom of the graph the 'true' reference parameter value, as found in the population, is pictured, with which each of the estimates needs to be compared. The results are illustrative for the other parameters. When five equally sized, large trials are randomly drawn from the same population (scenario 1), all methods lead to similar point estimates of the model parameters, but the BRE model has a much wider CI and a higher coverage. The difference in point-estimate between FFE and BFE is due to the different distributional assumptions: BFE assumes a binomial model, whereas FFE (implicitly) assumes a normal distribution.

In scenario 4 we added heterogeneity by drawing one of the trials from a less healthy population. The point estimates from the random effects (RE) models are further from the reference parameter. RE models assign a relatively greater weight to trials which outcomes differ from the rest. Due to the wider CIs, RE models are more likely to include the reference parameter value, but tend to overestimate uncertainty.

Varying trial sizes, with three trials from the same subgroup (scenario 7) leads to results comparable to scenario 4, where only one of the trials was drawn from this subgroup.

3.2 Model Parameters for 1,000 Repetitions

To investigate whether the results from the previous paragraph are due to chance, or if there are systematic differences, Table 4 presents a summary of the performance

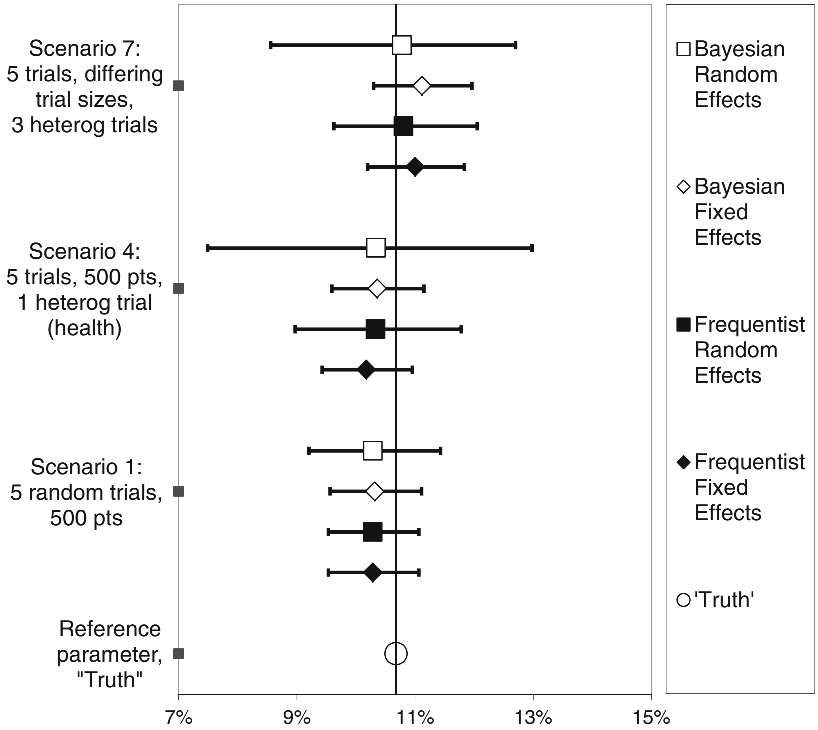

Fig. 3 Meta-analysis on the transition from the severe to very severe disease stage, for three of the seven scenarios, for the Usual Care arm, for one repetition. *pts* number of patients per trial, equal in each arm, *heterog* added heterogeneity by sampling from subpopulations

indicators over 1,000 repetitions. It reports the number of model parameters out of 33 for which the performance indicators are below or above certain threshold values. First we look at the I^2, averaged over 1,000 repetitions. For many parameters and scenarios, the mean of the I^2 statistic does not exceed 30 %, indicating no heterogeneity, even in scenarios where heterogeneity is built in. Some parameters show substantial heterogeneity, even if all trials are randomly drawn from the same population. The number of parameters with a mean I^2 below 30 % decreases when the amount of heterogeneity increases and the number of parameters with a mean I^2 above 50 % increases slightly.

When equally sized trials are randomly drawn from the same underlying population (scenario 1), the number of parameters with mean coverage below 90 % or above 98 % is comparable for FFE, FRE and BFE. BRE, on the other hand, shows no underestimation of uncertainty in any of the parameters, and an overestimation in 26 of the 33 parameters. FFE and BFE have a tendency to underestimate uncertainty when heterogeneity is added (scenarios 4 and 7), as is illustrated by the increasing number of parameters with a coverage lower than 90 %. It should be noted that an increase in bias and MAD also contributes to a lower coverage. In scenario 7, even the FRE model underestimates uncertainty for several parameters and the number of parameters where the uncertainty is overestimated decreases. BRE never underestimates uncertainty, and overestimates uncertainty for nearly all of the parameters in all scenarios. In fact, the coverage is 100 % in a large number of cases (not shown).

There are only small differences between methods in bias, with more bias in the scenarios with more added heterogeneity. There are, however, differences between the methods with respect to the MAD. The number of parameters where the MAD is larger than 5 % is smaller for the FFE and FRE, than for the BFE and BRE methods, regardless of heterogeneity. The BRE method generally yields point-estimates that are further away from the true population value than the other methods. Using RE models in scenarios with heterogeneity does not necessarily reduce bias. They may even increase bias, especially when the trials that differ from the others all differ in the same direction (scenario 7).

3.3 Health-Economic Outcomes for 1,000 Repetitions

Differences in model parameters may also lead to differences in HE outcomes. In Table 5, we show the mean HE outcomes over 1,000 repetitions, for both interventions and the difference between them. In scenario 1, all HE outcomes are very close to the true population value. In scenario 7, we can see that the point-estimates are further from the truth than is the case in the other two scenarios, for all methods of meta-analysis. On average, the number of QALYs estimated in each of the treatment arms is around 5 % below the true population value, and so is the difference in QALYs. In scenario 7, the fixed effects (FE) CIs (not shown) are comparable to those in scenario 1, but the RE CIs are much wider, especially for the BRE method.

Table 4 Summary of the result of meta-analysis on all parameters of the health-economic model. Means over 1,000 repetitions

Total number of parameters for which:	Scenario 1	Scenario 4	Scenario 7
Total number of parameters	33	33	33
Parameters influenced by added heterogeneity	0	24	24
Mean $I^2 < 30$ %: heterogeneity might not be important	27	27	22
Mean $I^2 > 50$ %: substantial heterogeneity	4	6	6
Mean coverage < 90 % (underestimation of uncertainty)			
FFE	6	9	23
FRE	6	0	21
BFE	6	9	23
BRE	0	0	0
Mean coverage > 98 % (overestimation of uncertainty)			
FFE	11	7	3
FRE	12	13	4
BFE	12	10	4
BRE	26	32	24
Mean bias > 2 %			
FFE	0	12	19
FRE	0	13	19
BFE	0	12	20
BRE	0	13	21
Mean MAD > 5 %			
FFE	0	3	16
FRE	0	5	16
BFE	3	6	17
BRE	5	9	17

BFE Bayesian fixed effects method, *BRE* Bayesian random effects method, *FFE* frequentist fixed effects method, *FRE* frequentist random effects method, *MAD* mean absolute deviation

Table 6 shows the coverage, bias and MAD for the difference between the two intervention groups. In general, we see that the coverage is larger in the RE methods, due to wider CIs which take heterogeneity into account. In addition, the Bayesian methods have higher coverage than the frequentist methods. Bias is generally low when no heterogeneity is included (scenario 1) and increases when it is (scenarios 4 and 7). The largest bias and MAD is found in the BRE method, for all outcomes in all scenarios. For the other three methods, bias and MAD are comparable. Despite the higher bias and MAD, the coverage of the BRE is still larger.

For the number of QALYs, events and total costs, statistical power (online appendix) is 100 % for all scenarios of FFE, FRE and BFE. It is slightly lower for the LYs for FFE, FRE and BFE, with a minimum of 96.7 %. For the BRE method, the statistical power for LYs ranges from 17.5 % (scenario 6) to 100 % (scenario 3). It is generally lower when there are more trials drawn from a subgroup of patients and when there is a difference in sample size between the trials.

Figure 4 shows the CE acceptability curves (CEAC) for scenario 7. The four graphs represent the four methods of meta-analysis. In each graph, we show the CEAC for ten repetitions, the median, 2.5th and 97.5th percentile over 1,000 repetitions. It is clear that even in this scenario with a lot of heterogeneity, the graphs are very similar for FFE, FRE and BFE. At a ceiling ratio of € 30,000 per QALY, which is very close to the true population ICER of € 28,020 (dashed vertical line), the median probability of New Intervention being cost effective is between 60–70 %, for these three methods. At a ceiling ratio of € 21,000, the median probability for all three methods is below 20 % and the 97.5th percentile is below 30 %. At € 39,000, the median probability is above 95 % and the 2.5th percentile is above 65 %, again for all three methods. Therefore, no great difference in policy decision would arise from using these three different methods of meta-analysis.

However, using BRE, the outcome would be different. Even at a ceiling ratio of € 48,000, the 97.5th percentile is below 60 %, and the median probability is below 90 %. Using BRE, a policy maker would be much less certain of the cost-effectiveness of the new intervention.

4 Discussion

In this study, we compared four methods of meta-analysis. Using a simulation study we could compare the HE outcomes to a gold standard and judge their statistical performance. In order to do this, we made a few crucial

Table 5 Health economic outcomes for three of the seven scenarios, both intervention arms and the difference. Means and range from the 2.5th and 97.5th percentiles over 1,000 repetitions

Scenario	Scenario 1 Five randomly drawn, equally sized trials			Scenario 4 Five equally sized trials; one trial drawn from a less healthy population			Scenario 7 Five equally sized trials; three trials drawn from a less healthy population		
Intervention arm	New Int	Usual	Diff	New Int	Usual	Diff	New Int	Usual	Diff
Number of QALYs									
Truth	0.540	0.485	0.054	0.540	0.485	0.054	0.540	0.485	0.054
FFE	0.542	0.488	0.054	0.533	0.480	0.053	0.515	0.464	0.051
FRE	0.541	0.487	0.054	0.532	0.479	0.053	0.514	0.463	0.051
BFE	0.541	0.486	0.054	0.531	0.478	0.053	0.513	0.461	0.052
BRE	0.540	0.487	0.054	0.531	0.478	0.052	0.513	0.462	0.051
Number of LYs									
Truth	0.786	0.740	0.046	0.786	0.740	0.046	0.786	0.740	0.046
FFE	0.789	0.744	0.045	0.781	0.738	0.044	0.767	0.723	0.043
FRE	0.788	0.744	0.045	0.780	0.736	0.044	0.766	0.723	0.044
BFE	0.787	0.742	0.045	0.779	0.735	0.044	0.764	0.720	0.044
BRE	0.787	0.743	0.045	0.779	0.735	0.043	0.764	0.721	0.042
Total costs (€)									
Truth	7,633	6,116	1,517	7,633	6,116	1,517	7,633	6,116	1,517
FFE	7,657	6,140	1,517	7,652	6,158	1,494	7,643	6,167	1,476
FRE	7,653	6,137	1,515	7,644	6,152	1,492	7,639	6,164	1,475
BFE	7,639	6,126	1,513	7,627	6,136	1,490	7,615	6,139	1,476
BRE	7,650	6,129	1,522	7,635	6,145	1,490	7,627	6,157	1,470

BFE Bayesian fixed effects method, *BRE* Bayesian random effects method, *Diff* difference between two intervention arms, *FFE* frequentist fixed effects method, *FRE* frequentist random effects method, *LY* life years, *New Int* new intervention, *QALYs* quality-adjusted LY, *Usual* usual care

Table 6 Health economic outcomes for three of the seven scenarios. Means of coverage, bias and mean absolute deviance (*MAD*) of the difference between two interventions, over 1,000 repetitions

Scenario	Scenario 1 Five randomly drawn, equally sized trials			Scenario 4 Five equally sized trials; one trial drawn from a less healthy population			Scenario 7 Five equally sized trials; three trials drawn from a less healthy population		
	Coverage (%)	Bias (%)	MAD (%)	Coverage (%)	Bias (%)	MAD (%)	Coverage (%)	Bias (%)	MAD (%)
Number of QALYs									
FFE	98.1	−0.2	8.3	97.4	−2.4	8.5	96.0	−5.1	9.5
FRE	98.8	−0.1	8.3	99.5	−2.2	8.5	99.3	−5.0	9.7
BFE	98.3	0.2	8.7	98.6	−2.1	8.8	97.9	−5.0	9.4
BRE	100.0	−1.0	9.6	100.0	−3.3	10.7	100.0	−6.6	13.4
Number of LYs									
FFE	98.2	−1.6	13.9	97.1	−3.8	14.2	97.5	−4.6	15.0
FRE	99.3	−1.4	14.0	99.1	−3.4	14.3	98.9	−4.4	15.5
BFE	98.6	−0.9	14.5	98.4	−3.1	14.8	99.2	−4.1	15.0
BRE	100.0	−2.3	16.1	100.0	−4.9	17.5	100.0	−6.9	20.8
Total costs									
FFE	98.5	0.0	5.1	98.2	−1.5	5.5	97.1	−2.7	5.9
FRE	99.3	−0.1	5.2	99.5	−1.7	5.6	99.4	−2.8	6.1
BFE	98.5	−0.3	5.3	98.5	−1.8	5.7	98.4	−3.2	6.1
BRE	100.0	0.3	6.4	100.0	−1.8	6.8	100.0	−3.1	8.1

BFE Bayesian fixed effects method, *BRE* Bayesian random effects method, *FFE* frequentist fixed effects method, *FRE* frequentist random effects method, *LY* life years, *QALYs* quality-adjusted LY

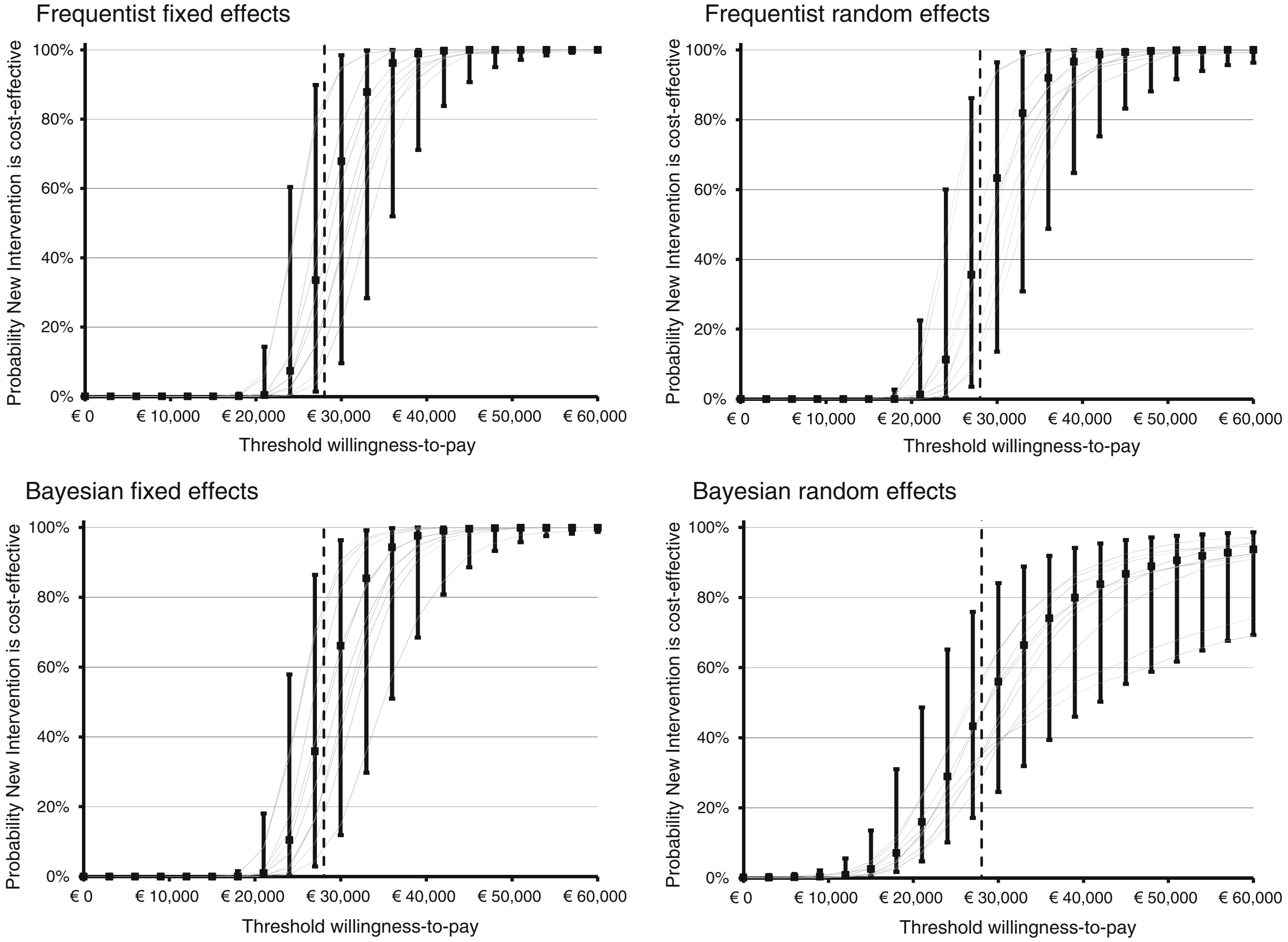

Fig. 4 Cost-effectiveness acceptability curves (*CEACs*) for the four models in the heterogeneous scenario 7. Graphs depicts median, 2.5th and 97.5th percentile CEACs over 1,000 repetitions, as well as the CEACs for the first 10 repetitions; *vertical, dashed line* is the 'true' population incremental cost-effectiveness ratio (*ICER*)

assumptions. First, we assumed that the researcher wants to estimate the parameter values of the entire population, not a subpopulation. This allows us to compare the results to the outcomes for the entire population. We also assumed the researcher was unaware of the fact that any heterogeneity was caused by sampling from subgroups. A researcher might not have combined the trials at all, had they been aware of the differences, by seeing the patient characteristics or trial protocols. A researcher might also have tried to compensate using regression methods, which were not the focus of the paper, nor would they be feasible with only five to ten trials.

With almost no heterogeneity, we found that the results of the FFE, FRE and BFE methods were comparable. With heterogeneity added to the trials, we saw differences on a parameter level, but these did not translate into important differences in HE outcomes. That could be because the HE model combines all parameter estimates and their uncertainties into one estimate of QALYs and total costs. All these uncertainties together may hide the (subtle) differences we have seen between the methods. In addition, we did not take structural uncertainty into account, which may exceed any parameter uncertainty.

Using any of these three methods would not lead to differences in policy decisions. Using BRE would, as it has a tendency to overestimate uncertainty and yield a larger probability that a new treatment is rejected or that more research is asked for where it might not be necessary. Partly, this is due to the number of trials included in the meta-analysis. Generally speaking, sophisticated methods, such as RE, require more data than simple methods, because of the increased number of parameters. This is particularly important for BRE, as it estimates between-study heterogeneity and also takes the uncertainty around this estimate into account. This can be estimated more precisely from ten trials than from five. In scenario 3, where we have the same amount of patients in ten trials instead of five, we have seen that the CI around the BRE is still larger than those of the other three methods, but the difference is much smaller. We also saw that the coverage for the BRE is much closer to 95 % and that uncertainty is overestimated in fewer parameters.

We speculate that with more than ten trials the differences might be even less pronounced and the BRE method will yield almost the same results as the other three methods, although the amount of uncertainty will always exceed that of the other methods. We did not test this assumption as this situation is unlikely within the scope of the expensive drug programme. In addition, time and budget constraints did not permit the calculation time needed for a simulation of this many trials, especially in a number of different scenarios.

Based on this, we recommend not using the BRE when only few sources of evidence are available. Unfortunately, this is more the rule than the exception, especially in the expensive drugs programme which was the reason to initiate this study. With only a few differences between the other three methods, we would personally favor FRE, as it automatically reduces to FFE in the absence of heterogeneity, is easy to implement and is more easily understood by physicians and policy makers who will be using the results.

By calculating outcomes for a number of scenarios, we have covered many of the different situations that are likely to arise in meta-analysis. We have drawn a few larger trials, but more smaller trials, and trials with differences in trial sizes. We have drawn trials randomly from the same population, one trial from a subgroup of patients, several trials from different subgroups and several trials from the same subgroup. Because of this, we feel that the results of our study are generalizable to other studies that use meta-analysis to obtain pooled estimates of parameters to fill a HE model.

We have made sure that the difference between the two interventions is large. When two interventions are much closer to each other, it unlikely this will change our conclusions regarding the methods of meta-analysis. The same is true for a longer time horizon, or including discounting.

Despite our feelings that the results are generalizable to other situations, there are several limitations to our study. The first limitation is that we have assumed that all data comes from the same set of trials. In practice, the data for transition probabilities will likely come from different sources than, for example, the RR for those transition probabilities or the utilities. The exact source of the evidence will not have an impact on the performance of the methods of meta-analysis. Therefore, we decided not to explore this extra complexity in this paper.

Another limitation is the choice of prior for the Bayesian models. The use and choice of priors is an important subject when discussing the Bayesian methodology. Any Bayesian calculation can be affected by the type of priors used. In the case of meta-analysis, a small number of studies is extra vulnerable to the type of prior [8, 17]. As we did not assume the researcher to have prior information, we also used so called vague, or flat priors. Even though they are supposed to be ‘uninformative’, they may influence the outcomes, especially the posterior scale parameters [17]. We tested several different specifications of the priors but did not find any differences in outcomes, likely from the relative simplicity of the models used. However, researchers using the BFE or BRE should keep these restrictions in mind and different priors may lead to different results.

Our results are not generalizable to network meta-analysis and should only be used in the case of a pair-wise comparison of two interventions. In the case that more than two comparators are available, other methods of meta-analysis are available, which make use of all the available evidence [18–21].

We have seen that both the RE methods and the appropriate measure for heterogeneity, I^2, have a tendency to detect heterogeneity, when trials have differences in number of patients, even with a large number of total patients, randomly drawn from the same underlying population. This is a very common occurrence in meta-analysis and may lead to too conservative CIs as none of the methods can make the distinction between sampling error and heterogeneity. Trials can therefore be considered heterogeneous, not only when one or more trials are drawn from a (different) subgroup of patients, but also when all trials are randomly drawn from the same population, but with differences in trial sizes. At the same time, with heterogeneity built in, many of the parameters show no important degree of heterogeneity. From this we can see that the I^2 might be an imperfect measure for heterogeneity, at least with a relatively low number of trials.

In our simulation study, we have made sure that the reference parameters are not close to their natural limits; for example, probabilities or costs close to 0. In cases when the reference parameters are closer to these limits, we expect that the Bayesian methods will have model parameters that are closer to the true population value than the frequentist methods. First of all, frequentist methods usually need a correction term (continuity correction) if one of the trial parameters is 0, because it will not be possible to calculate the necessary standard errors otherwise. Bayesian methods do not. In addition, Bayesian methods may use a bounded likelihood function, while frequentist methods always implicitly use a normal distribution. This might be a valid reason to prefer Bayesian methods over frequentist methods.

The transition probabilities and probabilities to experience an event in the New Intervention arm were calculated using the model parameter in the Usual Care arm, and the corresponding RR. Results using the risk difference were similar and therefore not shown.

In many HE models, many input parameters need to be estimated. When more than one input parameter is estimated from the same set of sources, we recommend heterogeneity is not checked for each parameter separately, but rather for the set of trials. If statistics indicate trials are homogeneous for one parameter, but heterogeneous for another, it is recommended that all parameters are calculated using the same type of model. The model type selection should be based on *trial heterogeneity* rather than *parameter heterogeneity*.

5 Conclusion

In conclusion, the FFE, FRE and BFE meta-analysis methods led to comparable HE outcomes, even in scenarios where we built in heterogeneity. The differences that we see between the methods point towards a broader CI (which is translated in a higher coverage), a higher MAD and a lower statistical power for Bayesian methods compared with frequentist methods, and for RE methods compared with FE methods. RE methods do not necessarily reduce bias when heterogeneity is added to the trials, and may even increase bias in certain situations. BRE tends to overestimate uncertainty reflected in the shape of the CEAC. Based on this study, we recommend the FRE method as the preferred method of meta-analysis.

Acknowledgments The authors wish to acknowledge the support of Prof. Lidia Arends, Ph.D. and Prof. Jos Kleijnen, Ph.D. for their valuable contributions to the design and execution of this study. We also thank the Netherlands Organization for Health Research and Development (ZonMW), which sponsored this study under project number 152002002. The authors would also like to thank four anonymous reviewers for their valuable comments on the final manuscript.

Author Contributions The last three authors initiated and designed the study. The first author performed the simulation study and wrote the manuscript. The last three authors contributed to writing, reviewing and approving the manuscript. All authors contributed to analyzing and interpreting the outcomes of the study.

Competing Interests PV has no competing interests. MA has no competing interests. MO has no competing interests. MR has no competing interests.

References

1. NZa. Beleidsregel Dure Geneesmiddelen [Policy rule Expensive Drugs] (BR-CU-2017). 2011. http://www.nza.nl/regelgeving/beleidsregels/ziekenhuiszorg/BR-CU-2017. Accessed 15 June 2011.
2. Cooper NJ, Sutton AJ, Abrams KR, Turner D, Wailoo A. Comprehensive decision analytical modelling in economic evaluation: a Bayesian approach. Health Econ. 2004;13(3):203–26.
3. Ades AE, Sculpher M, Sutton A, Abrams K, Cooper N, Welton N, et al. Bayesian methods for evidence synthesis in cost-effectiveness analysis. Pharmacoeconomics. 2006;24(1):1–19.
4. Riley RD, Simmonds MC, Look MP. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. J Clin Epidemiol. 2007;60(5):431–9.
5. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med. 2001;20(6):825–40.
6. Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. 2001;10(4):277–303.
7. Sutton AJ, Higgins JP. Recent developments in meta-analysis. Stat Med. 2008;27(5):625–50.
8. Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions 5.0.2, updated September 2009. Available at http://www.cochrane-handbook.org/.
9. Oppe M, Al M, Rutten-van Molken M. Comparing methods of data synthesis: re-estimating parameters of an existing probabilistic cost-effectiveness model. Pharmacoeconomics. 2011;29(3): 239–50.
10. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557–60.
11. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177–88.
12. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539–58.
13. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. London: Chapman & Hall; 1995.
14. Carlin BP, Louis TA. Bayes and empirical Bayes methods for data analysis. London: Chapman & Hall; 1996.
15. Jaynes E. Confidence intervals vs Bayesian intervals. In: Harper W, Hooker CA, editors. Foundations of probability theory, statistical inference, and statistical theories of science. Dordrecht: D Reidel; 1976. p. 175.
16. O'Hagan A, Luce B. A primer on Bayesian statistics in health economics and outcomes research. Sheffield: Centre for Bayesian Statistics in Health Economics; 2003.
17. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Stat Med. 2005;24(15):2401–28.
18. Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. BMJ. 2003;326(7387):472.
19. Strassmann R, Bausch B, Spaar A, Kleijnen J, Braendli O, Puhan MA. Smoking cessation interventions in COPD: a network meta-analysis of randomised trials. Eur Respir J. 2009;34(3):634–40.
20. Puhan MA, Bachmann LM, Kleijnen J, Ter Riet G, Kessels AG. Inhaled drugs to reduce exacerbations in patients with chronic obstructive pulmonary disease: a network meta-analysis. BMC Med. 2009;14(7):2.
21. Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. J Am Stat Assoc. 2006;101:447–59.