Original Article

Medical Decision Making
2018, Vol. 38(4) 531–542

Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0272989X18759488
journals.sagepub.com/home/mdm

Incorporating Genuine Prior Information about Between-Study Heterogeneity in Random Effects Pairwise and Network Meta-analyses

Shijie Ren, Jeremy E. Oakley, and John W. Stevens

Abstract

Background. Pairwise and network meta-analyses using fixed effect and random effects models are commonly applied to synthesize evidence from randomized controlled trials. The models differ in their assumptions and the interpretation of the results. The model choice depends on the objective of the analysis and knowledge of the included studies. Fixed effect models are often used because there are too few studies with which to estimate the between-study SD from the data alone. **Objectives.** The aim of this study was to propose a framework for eliciting an informative prior distribution for the between-study SD in a Bayesian random effects meta-analysis model to genuinely represent heterogeneity when data are sparse. **Methods.** We developed an elicitation method using external information, such as empirical evidence and expert beliefs, on the "range" of treatment effects to infer the prior distribution for the between-study SD. We also developed the method to be implemented in R. **Results.** The 3-stage elicitation approach allows uncertainty to be represented by a genuine prior distribution to avoid making misleading inferences. It is flexible to what judgments an expert can provide and is applicable to all types of outcome measures for which a treatment effect can be constructed on an additive scale. **Conclusions.** The choice between using a fixed effect or random effects meta-analysis model depends on the inferences required and not on the number of available studies. Our elicitation framework captures external evidence about heterogeneity and overcomes the assumption that studies are estimating the same treatment effect, thereby improving the quality of inferences in decision making.

Keywords

few studies, health technology assessment, heterogeneity, meta-analysis, prior elicitation, random effects

Date received: August 3, 2017; accepted: January 4, 2018

Evidence of clinical effectiveness can arise from multiple sources. Pairwise meta-analysis (MA) is an established statistical tool for estimating the relative efficacy of 2 interventions evaluated in randomized controlled trials. In the absence of head-to-head studies, network meta-analysis (NMA) can be used to synthesize all available evidence and make simultaneous comparisons between treatments.

A pairwise MA and NMA can be conducted using a fixed effect or a random effects model. These models differ in their assumptions as well as in the interpretation of

School of Health and Related Research, University of Sheffield, Sheffield, England, UK (SR, JWS); and School of Mathematics and Statistics, University of Sheffield, Sheffield, England, UK (JEO). The work was done at School of Health and Related Research and School of Mathematics and Statistics in University of Sheffield. The author(s) received no financial support for the research, authorship, and/or publication of this article. The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Corresponding Author
Shijie Ren, Health Economics and Decision Science (HEDS), School of Health and Related Research, University of Sheffield, Regent Ct, 30 Regent St, Sheffield, England S1 4DA, UK; +44 (114) 222 0696. (s.ren@sheffield.ac.uk)

the treatment effects.[1–4] The choice of which model to use depends on the objective of the analysis and knowledge of the included studies. In this paper, we investigate the circumstances when and rationale for using these 2 models in single technology appraisals (STAs) submitted to the National Institute for Health and Care Excellence (NICE), which appraises health technologies and provides guidance to the National Health Service in England. We also propose how to overcome the problem of imprecise estimates of the heterogeneity parameter in the absence of sufficient sample data.

A fixed effect model would be appropriate if the objective is to determine whether the treatment had an effect in the observed studies (i.e., a conditional inference) or it would be appropriate when it is believed that the true treatment effects in each study are the same. Heterogeneity is expected in MAs because they combine studies that have clinical and methodological heterogeneity.[5] A random effects model would be preferred because it allows for heterogeneity in the treatment effects among the studies and allows the results to be generalized beyond the studies included in the analysis. Nevertheless, fixed effect models are still commonly used even when heterogeneity is expected.

Parameters can be estimated from either a frequentist or Bayesian perspective. The Bayesian approach provides more natural and useful inference, can incorporate external information, and is ideal for problems of decision making. There has been an increase in the use of Bayesian evidence synthesis in submissions to NICE, perhaps primarily because the evidence synthesis Technical Support Documents (TSDs) issued by the NICE Decision Support Unit (DSU)[6–11] advocate the Bayesian approach.

A random effects model requires an estimate of the between-study SD. When the number of included studies is small, the estimate of the between-study SD will be highly imprecise and biased in a frequentist framework such as using DerSimonian and Laird estimate.[1] Similarly, a Bayesian analysis of only limited data, using a standard, vague or weakly informative prior distribution for the between-study SD will give implausible posterior distributions.[12] A proper Bayesian analysis requires genuine specification of the prior distribution using external evidence, typically including experts' beliefs. Note that a judgement that a posterior distribution is implausible by an individual suggests that he/she must have some prior beliefs to make elicitation feasible.

NICE DSU TSD[7] suggests comparing goodness-of-fit of both fixed effect and random effects models using the deviance information criteria (DIC).[13] However, when the number of studies is small, it is likely that either model would at least provide an adequate fit to the data, specifically when the data are not sufficiently informative to learn about the between-study SD. Rather than goodness-of-fit, the issue is therefore how best to appropriately represent uncertainty about the treatment effect. When heterogeneity is expected, a fixed effect model is likely to be overconfident, and a random effects model with a vague prior is likely to be underconfident: a compromise between these 2 extremes is needed, which can be achieved with a more informative prior distribution.

Higgins et al.[14] presented an example of a Bayesian MA of MAs to create a predictive distribution for the between-study variance in gastroenterology. Other authors have generated predicative distributions for the heterogeneity expected in future MAs in more general settings using data from the Cochrane Database of Systematic Reviews for a log odds ratio (LOR),[15–17] and a standardized mean difference.[18] Smith et al.[19] constructed an informative prior distribution for the between-study variance using a gamma distribution by assuming that odds ratios (ORs) between studies have roughly one order of magnitude spread, and that it is very unlikely that the variability in treatment effects between studies varies by 2 or more orders of magnitude. Spiegelhalter et al.[20] suggested that a half-Normal distribution could be used as a prior distribution for the between-study SD and showed how to interpret the prior distribution. NICE DSU TSD[8] also suggested using an informative half-Normal prior distribution with a mean of 0 and a variance of 0.32^2, representing the belief that 95% of the study-specific ORs lie within a factor of 2 from the median OR for each comparison. Both half-Normal prior distributions are proposed for treatment effects measured by ORs. To the best of our knowledge, there has been little work on the formal elicitation of experts' beliefs for the between-study SD in random effects MA models.

To investigate the application of fixed effect and random effects models in submissions to NICE, we conducted a review of all the STAs completed up to 31 October, 2016. (Although this is a selective set, we believe the findings likely to be consistent with the rationale for analyses by international pharmaceutical companies to other HTA decision makers.) The results of the review are presented in Section 2. In Section 3, we propose novel methods to construct an informative prior distribution for the between-study SD using external information for all common types of outcome measures. Examples of re-analyzing 2 STAs using the proposed elicitation framework are given in Section 4.

A Review of NICE STAs

Two hundred and thirty-nine NICE STAs were completed between September 2005 (when the STA process was introduced) and 31 October 2016. After assessment by SR, a final set of 183 STAs was identified for review. Figure 1 presents a flow chart of the identification, inclusion, and exclusion of STAs. We have only reviewed the original companies' submissions, and not considered additional analyses that may have occurred during the appraisal process.

Thirty-eight STA submissions used pairwise MAs with a single approach being applied within each submission: 25 (66%) used a frequentist approach to estimate parameters and make inferences; 8 (21%) pooled individual patient-level data across studies; and, in 5 cases (13%), it was unclear which method was used. Ninety-three STA submissions included NMAs (multiple approaches may have been used in one submission): 71 (76%) of these used either a Bayesian or a frequentist NMA, 41 (44%) used Bucher indirect comparisons[21], and 1 (1%) didn't report the method.

We extracted the rationale for using fixed effect and random effects model for both pairwise MAs and NMAs. The findings of the review are presented in Table 1 and are summarized as follows:

- All submissions that performed pairwise MAs used a frequentist approach, and most that performed NMAs used a Bayesian approach (90%).
- 71% of the submissions that performed fixed effect pairwise MAs did not provide a justification for the model choice. For the submissions that performed random effects MAs, 60% gave no justification for the model choice. Fewer submissions using NMAs provided no justification for the model choice: 25% and 27% for fixed effect and random effects model, respectively.
- The most frequently stated reason for the use of a fixed effect model was that there were too few studies to conduct a random effects model.
- In some cases, where heterogeneity was noted, there was an acknowledgement that a random effects model would be appropriate but it was not used when there were only few studies.
- Among the pairwise MAs that used a frequentist approach, the choice of fixed effect or random effects model was typically assessed using the Q-statistic/I^2-statistic.
- When Bayesian fixed effect and random effects models were both used in a submission, the most popular method for choosing the final model was comparing the DIC statistic for the 2 models (62%).
- Providing either a fixed effect or random effects model within a sensitivity analysis was observed in both pairwise MAs (9%) and NMAs (21%) in the case where both models were used.
- Four submissions performed sensitivity analyses using different prior distributions for the between-study SD. TA288[22] considered the possibility of using alternative data sources to inform the prior distribution but concluded that no suitable sources were available. TA341[23] used a prior distribution informed using predictive distributions proposed by Turner et al.[16] TA173[24] used a half-Normal prior based on a re-analysis of the data from a previous systematic review. TA343[25] used a half-Normal prior distribution suggested by NICE DSU TSD.[8]

Overall, we found that the most frequently stated reason for the use of a fixed effect model was that there were too few studies to conduct a random effects model but not that there was unlikely to be heterogeneity or that a conditional inference was of interest. This showed that there is a need for more guidance on properly accounting for heterogeneity when the number of included studies is small. We now present a framework for constructing prior distributions for the heterogeneity parameter using external information such as empirical evidence and experts' beliefs.

General Elicitation Framework

For simplicity, we suppose that there is one female expert and the elicitation is conducted by a male facilitator. We do not consider issues such as the selection and training of experts, motivation, and how to elicit a prior distribution from multiple experts, which are covered elsewhere.[26–28] The general elicitation framework proposed in this section is for performing a pairwise MA. An extension to the approach for use in NMAs is discussed later.

We envisage that the elicitation will take place after specification of the decision problem and completion of the systematic literature review, with the finding that few studies satisfy the inclusion criteria for the MA. The expert making the judgments could be a clinician or an analyst who conducts the MA. We envisage her to have expertise specific to the disease area and treatments under investigation. She will be given the information on the decision problem, including population, intervention, control, outcome, and the summary of

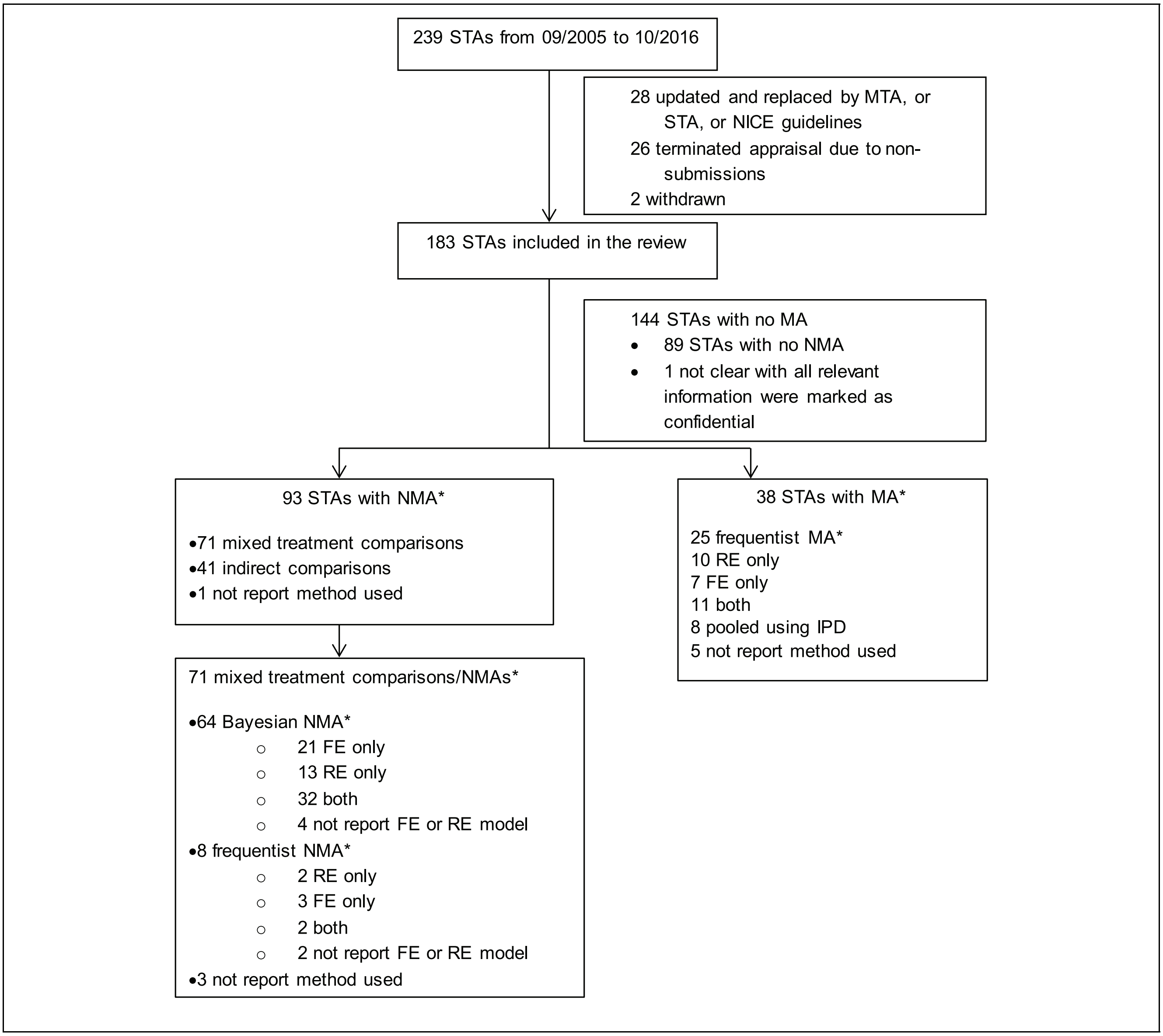
239 STAs from 09/2005 to 10/2016
28 updated and replaced by MTA, or STA, or NICE guidelines
26 terminated appraisal due to non-submissions
2 withdrawn
183 STAs included in the review
144 STAs with no MA
• 89 STAs with no NMA
• 1 not clear with all relevant information were marked as confidential
93 STAs with NMA*
•71 mixed treatment comparisons
•41 indirect comparisons
•1 not report method used
38 STAs with MA*
25 frequentist MA*
10 RE only
7 FE only
11 both
8 pooled using IPD
5 not report method used
71 mixed treatment comparisons/NMAs*
•64 Bayesian NMA*
○ 21 FE only
○ 13 RE only
○ 32 both
○ 4 not report FE or RE model
•8 frequentist NMA*
○ 2 RE only
○ 3 FE only
○ 2 both
○ 2 not report FE or RE model
•3 not report method used

Figure 1 Flow chart showing the identification, inclusion, and exclusion of reviews. STA, single technology appraisal; MTA, multiple technology appraisal; RE, random effects; FE, fixed effect; MA, meta-analysis; NMA, network meta-analysis; IPD, individual patient-level data. *Multiple analyses and analyses for multiple outcomes may have been conducted in one submission.

baseline characteristics of the included studies, and is encouraged to think about any potential treatment effect modifiers.

Suppose that there are S studies included in the MA, and that the treatment effect in study i is denoted by δ_i, for $i = 1 \ldots S$, expressed on some appropriate additive scale. The expert is required to make judgements about the likely variability in $\delta_1, \ldots, \delta_S$ between studies. For any 2 studies i and j, one could make judgements about the relative treatment effect δ_i/δ_j, i.e., “the treatment effect in one study could be x times that of the treatment effect in another.” Alternatively, one could consider the difference in treatment effects $\delta_i - \delta_j$, i.e., “the treatment effect in one study could exceed that in another study by x units’. In this paper, we consider the former case only, building on the discussion and analysis by previous authors.[19,20] In the latter case, elicitation methods for variances discussed by Alhussain and Oakley[29] could be considered.

Table 1 Justifications of Model Choice in Submissions

Method Used (Number of Submissions)		Justification	N (%)
Pairwise meta-analysis (38[a])	Fixed effect model only (7)	No justification	5 (71%)
		Check heterogeneity using test statistic	2 (29%)
	Random effects model only (10)	No justification	6 (60%)
		Allow for heterogeneity	3 (30%)
		Check heterogeneity using test statistic	1 (10%)
	Both models (11[b])	Not clear which model was a base case	5 (45%)
		Check heterogeneity using test statistic	4 (36%)
		One model as sensitivity analysis	1 (9%)
		Checking inclusion criteria	1 (9%)
	Pooling using individual patient-level data (8)		
	Unclear (5)		
Network meta-analysis (71[a])	Fixed effect model only (24)	Insufficient data	17 (71%)
		No justification	6 (25%)
		Check heterogeneity using test statistic	1 (4%)
	Random effects model only (15)	Allow for heterogeneity	4 (27%)
		No justification	4 (27%)
		Sufficient data	2 (13%)
		Check heterogeneity using test statistic	1 (7%)
		Same model as a previous study	1 (7%)
		Count for correlations	1 (7%)
		Count for multi-arms	1 (7%)
		Unclear	1 (7%)
	Both models (34[b])	Based on deviance information criteria	21 (62%)
		One model as sensitivity analysis	7 (21%)
		Final model fixed effect because of insufficient data	4 (12%)
		Not clear which model was a base case	3 (9%)
		Compare the credible intervals	1 (3%)
		Presence of closed loops	1 (3%)
	Unclear (2)		

[a] Multiple analyses and analyses for multiple outcomes may have been conducted in one submission.
[b] Multiple reasons for model choice may have been used in one analysis.

We assume that $\delta_1, \ldots, \delta_S \sim N(d, \tau^2)$, where d is the average treatment effect in a population of treatment effects and τ^2 is the between-study variance, which represents the heterogeneity in treatment effects between studies. This is the standard model when δ_i is a LOR, log hazard ratio, or mean difference.[7] We define R to be the ratio of the 97.5th percentile to the 2.5th percentile of treatment effects on the natural scale from a population of treatment effects. When the additive treatment effect estimated in the MA is on the log scale, we propose to elicit the treatment effect on the natural scale. For example, if δ_i is a LOR, then we propose using OR_i in the elicitation, which is $\exp(\delta_i)$, and $R = OR_{97.5}/OR_{2.5}$ is the ratio of the 97.5th percentile to the 2.5th percentile of ORs in a population of treatment effects, roughly representing the "range" of ORs. Noting that $\log R = \delta_{0.975} - \delta_{0.025}$, we link R to τ via

$$\begin{aligned} \delta_{97.5} - \delta_{2.5} &= 2 \times 1.96\tau = 3.92\tau \\ \Rightarrow \log R &= 3.92\tau \\ \Rightarrow \tau &= \frac{\log(R)}{3.92} \end{aligned} \qquad (1)$$

We propose asking the expert to make judgements about R, from which judgements about τ can be inferred using equation (1). However, given the somewhat abstract nature of R, we suggest providing the less formal definition to the expert: she is asked to consider the ratio of the largest to the smallest treatment effect on the natural scale that could arise over a set of studies (though the expert should be told that "largest" and "smallest" will be interpreted as 97.5th and 2.5th percentiles), ignoring sampling variability within studies.

A Three-stage Procedure for Eliciting the Prior Distribution for the Between-study SD

In some cases, even with adequate training, the expert may find it difficult to make the judgements about R that are necessary to obtain a distribution for τ. Consequently, we propose a 3-stage procedure depending on the judgements that the expert can make. We firstly present this procedure when the treatment effect is an LOR, and then discuss modifications for treatment effects reported on different scales. Instructions for implementing our method using the R package SHELF[30] are given in Appendix 3.

Stage 1: Confirmation of the Need for a Random Effects Model. The fixed effect model is a special case of the random effects model, corresponding to the judgement that $P(R = 1) = 1$, i.e., the expert is certain that the largest OR is the same as the smallest OR in a set of studies. The expert is asked to either rule out or accept this case:

- "Can you be certain that the treatment effects across the studies will be identical, ignoring within-study sampling variability?"

If she is certain that this will be the case, then a fixed effect model should be used with appropriate justification provided. Otherwise, we proceed to Stage 2.

Stage 2: Consideration of an Upper Bound for R. If a random effects model is deemed to be appropriate, the expert is then asked if she can provide an upper bound for R. She is asked:

- "Let R be the ratio of the largest to the smallest OR. Are you able to judge a maximum plausible value for R? Denoting this limit by R_{max}, this means that you would think values of R above R_{max} are too implausible to be contemplated."

If the expert's answer for R_{max} is, for example, 10, this means that she believes that the OR in one study could be no more than 10 times that of the OR in another, i.e. one order of magnitude. If the expert is not able to provide a value R_{max}, then we recommend using the prior distribution proposed by other author.[15–17] For example, Turner et al.[16] proposed a prior distribution for the between-study variance in a general setting with the treatment effect measured by a LOR,

$$\log \tau^2 \sim N\left(-2.56, 1.74^2\right). \tag{2}$$

If she can provide R_{max}, then we proceed to Stage 3. Note that we do not propose asking for a lower limit for R because we think experts would typically not want to rule out the case $R = 1$ as impossible. The expert could also provide a lower limit R_{min}, if she wished, with R_{min} replacing the lower limit of 1 in the following.

Stage 3: Consideration of a Full Distribution for R. We now ask if the expert judges some values in the range $[1, R_{max}]$ to be more likely than others, and if she is able to express her beliefs using the roulette elicitation method.[31] If she is not able to make such judgements, then we propose using prior distributions proposed by others,[15–17] but now truncated to $\left[0, \left(\frac{\log(R_{max})}{3.92}\right)^2\right]$.

If she can continue with the roulette method, then the range $[1, R_{max}]$ is divided into several equal-width "bins." The expert is asked to specify her probability of R lying in a particular bin by placing "chips" in that bin, with the proportion of chips allocated representing her probability. The number of chips given to the expert is specified by the facilitator. For example, if in total 20 chips are used, then each chip represents a probability of 0.05. An illustration is given in Figure 2a. Here, the expert has chosen $R_{max} = 10$. By placing 5 chips out of 20 in the bin (2, 3], she has expressed a judgement that $P(2 < R \leq 3) = \frac{5}{20}$.

We suggest fitting either a gamma or lognormal distribution to the elicited probabilities by choosing the distribution parameters to minimize the sum of squares between the elicited and fitted cumulative probabilities. The R package SHELF[30] will identify the best fitting distribution out of the gamma and lognormal; although, there is unlikely to be much difference in the fitted distributions in most practical situations. Given that R has a lower limit of 1, the package will fit a gamma or lognormal distribution to $R - 1$. Hence, using a lognormal distribution, for example, we will have a prior for τ specified via

$$\log(R-1) \sim N(m, v),$$
$$\tau = \frac{\log R}{3.92},$$

where m and v are the mean and variance for the elicited lognormal distribution.

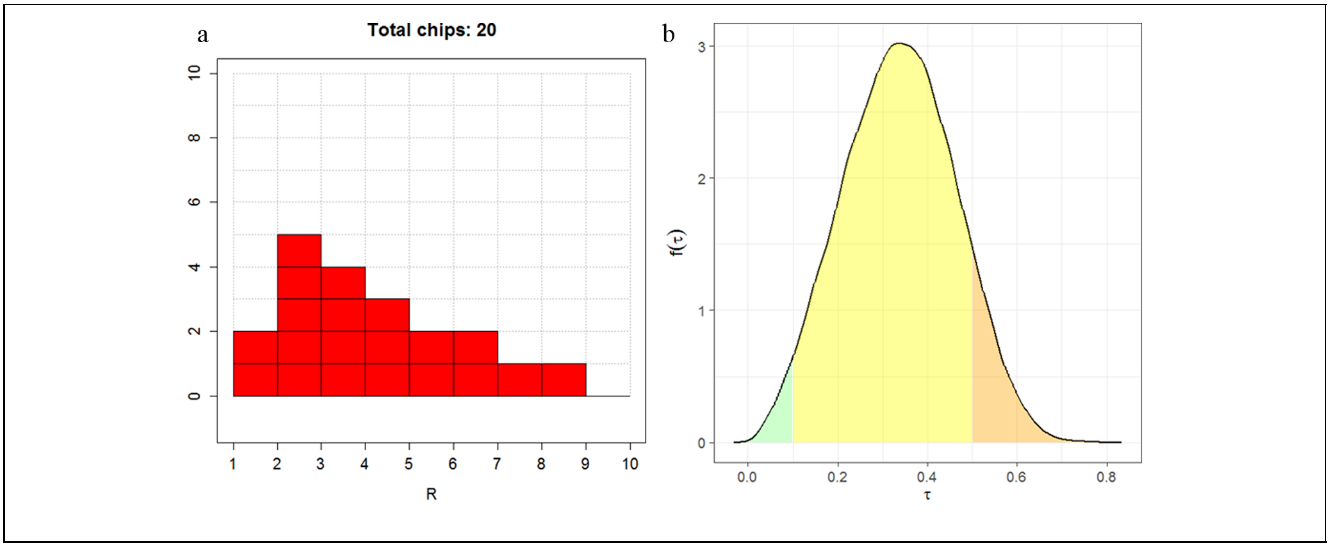

Figure 2 (a) Eliciting beliefs about R with the roulette method. (b) The implied distribution of τ, following the elicited judgements about R shown in (a). The probabilities of "low," "moderate," and "high" heterogeneity are in green, yellow, and orange, respectively (with negligible probability of 'extreme' heterogeneity).

Feedback

We propose providing feedback to the expert about the implied distribution of τ using the probabilities of implied low/moderate/high/extreme high heterogeneity, regardless at which stage in the above procedure the elicitation is concluded. Spiegelhalter et al.[20] suggested that values of τ between 0.1 and 0.5 are considered as reasonable heterogeneity in many contexts (what we describe as moderate), from 0.5 to 1.0 as fairly high heterogeneity (what we describe as high) and above 1.0 fairly extreme heterogeneity (what we describe as extreme). The probability of τ in the range of less than or equal to 0.1, (0.1, 0.5], (0.5, 1.0] and above 1.0 will be provided to the expert. The distribution can be displayed using a kernel density estimate or histogram of a large sample of randomly generated values of τ. An illustration using SHELF package[30] is given in Figure 2b. The probabilities of "low," "moderate," and "high" heterogeneity are approximately 0.03, 0.87, and 0.1, respectively (with negligible probability of "extreme" heterogeneity) given the elicited judgements about R in Figure 2a.

Other Types of Outcome Measures

Other scale-free outcome measures include hazard ratio, relative risk, and ratio of means for continuous outcomes.[32,33] The 3-stage procedure could be used in these cases; although, it is less clear that the prior distributions proposed by previous authors[15–17] would be appropriate because the distributions were derived based on empirical evidence of heterogeneity in ORs in MAs. It is likely that an elicitation exercise considering a full distribution for the ratio of treatment effects R (for example, the ratio of the largest to the smallest hazard ratio among the studies) would be required in these cases.

When the outcome measure is continuous or ordered categorical with the MA model using the identity or probit link functions, the expert may find it difficult to express beliefs about the "range" of treatment effects because the continuous measurement is not unit-free and the probit scale is difficult to interpret directly. We propose using the method described in Section 3.1 with the following modification:

1. Dichotomize the response using some appropriate cut-off c, to define a new treatment effect δ_i on the OR scale.
2. Considering ORs for the dichotomized response, use the 3-stage procedure to elicit a prior distribution for τ, the variability in LORs in a population of studies.
3. Given a prior distribution for τ, convert this to a prior distribution for the between-study SD $\tilde{\tau}^2$ on the original scale (i.e., probit or continuous) via
$$\tilde{\tau} = \omega\tau,$$
with $\omega = \frac{\sqrt{3}}{\pi}$ for the probit scale, and $\omega = \sigma\frac{\sqrt{3}}{\pi}$ for the continuous scale, where σ is an estimate of an individual level SD. The estimate could be a

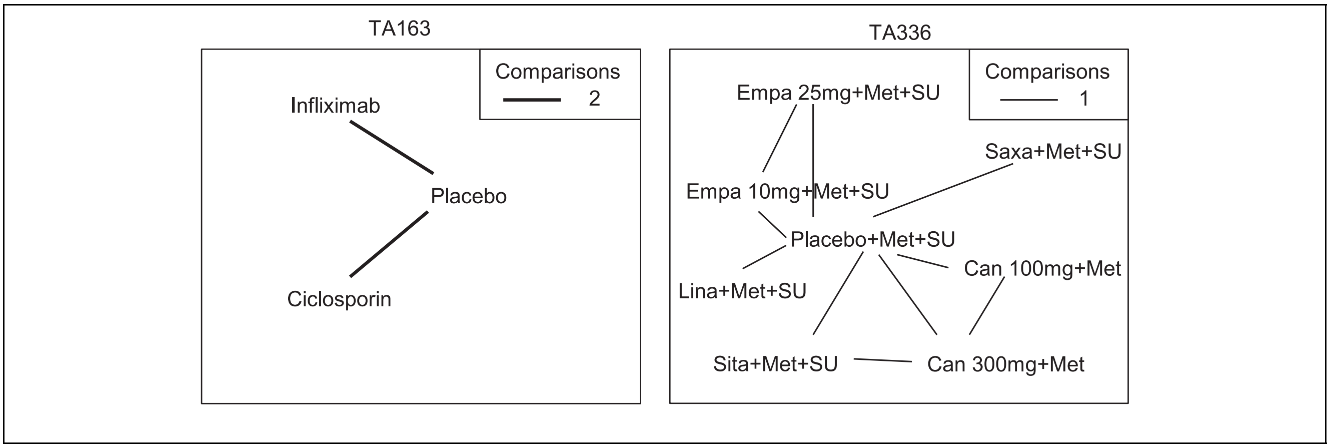

Figure 3 Network diagram for TA163[35] and TA336[36] used in the example. The thickness of the line represents the number of times pairs of treatment have been compared in studies. Empa, empagliflozin; Lina, linagliptin; Sita, sitagliptin; Saxa, saxagliptin; Can, canagliflozin; Met, metformin; SU, sulphonylurea.

summary measure of the SDs in the included studies, pooled from included studies, or be obtained from a single representative study.

Details of the derivation can be found in Appendix 1.

Network Meta-analysis

NMAs typically assume a homogeneous variance model.[7,14,34] A similar elicitation method as described above can be used to elicit the common heterogeneity parameter in an NMA. We suggest asking the expert for the "range" of treatment effects *R* for a pairwise comparison based on the one that the expert is most comfortable about in expressing her beliefs. When giving feedback to the expert on the probability that the heterogeneity would be low, moderate, high, or extremely high, we could ask the expert whether she would agree with these elicited probabilities for other pairwise comparisons in the network.

Examples: Reanalysis of Two STAs

We re-analyzed the data from 2 NICE STAs (TA163[35] and TA336[36]) to demonstrate the use of our proposed method. BUGS code incorporating the different prior specifications is provided in Appendix 2.

TA163[35] was a technology appraisal of infliximab for treating acute exacerbations in adults with severely active ulcerative colitis. Data were available from 4 studies of 3 treatments (placebo, infliximab and ciclosporin; Figure 3). The outcome measure was the colectomy rate at 3 mo. A fixed effect model was used in the original submission. Table 2 presents results from a Bayesian NMA using a fixed effect model, a random effects model with a vague prior distribution uniform [0, 5], as used in Dias et al.,[7] and 3 alternative informative prior distributions: the prior distribution in equation (2), both untruncated and truncated so that $R_{max} = 10$, and an elicited prior distribution using the proposed method in Section 3. As an illustration, the elicited judgements were those of the author (SR). Results are presented as the median with 95% credible and prediction intervals based on 40,000 iterations from the Markov chain after a burn-in of 60,000 iterations using the software OpenBUGS.[37]

As expected, the DIC statistics for the 5 models were fairly similar: 34.72, 33.44, 34.70, 35.19, and 34.60, and did not provide support for any one model over another. The fixed effect model showed that there was evidence that ciclosporin reduced the colectomy rate at 3 mo relative to placebo in the studies included in the NMA, whereas there was insufficient evidence to conclude that infliximab had an effect relative to placebo in the included studies. As expected, the results of the random effects model demonstrated the sensitivity of the results to the different prior beliefs about the heterogeneity parameter. The uniform [0, 5] prior for the heterogeneity parameter was not "updated" appreciably by the data (Figure 4) and gave very different results compared to the fixed effect model (Table 2). There was a large posterior probability, 0.87, that heterogeneity was extremely high, equivalent to saying that the probability that the OR in a study could be 50 or more times that of the OR in another was 0.87 (The interpretation of the heterogeneity parameter can be found in Appendix 1). This is

Table 2 Comparison of Results Obtained from Fixed Effect and Random Effects Models[a]

Example 1: TA163[35]. Colectomy rate at 3 mo: treatment effect on log odds ratio scale	**OR, median (95% CrI) ciclosporin v. placebo**	**OR, median (95% CrI) infliximab v. placebo**	P_L	P_M	P_H	P_{EH}
FE	0.13 (0.03 to 0.44)	0.72 (0.18 to 2.70)	0	0	0	0
RE with $\tau_{OR} \sim uniform[0, 5]$	0.02 (0 to 1.46) **0.03 (0** to **33.02)**	0.70 (0.01 to 84.59) **0.69 (0** to **2498.82)**	0.01	0.05	0.07	0.87
RE with $\tau^2_{OR} \sim lognormal(0.256, 1.74^2)$	0.11 (0.01 to 0.48) **0.12 (0.01** to **0.62)**	0.71 (0.14 to 3.25) **0.71 (0.10** to **4.83)**	0.11	0.62	0.18	0.08
RE with $\tau^2_{OR} \sim lognormal(0.256, 1.74^2)$ truncated with upper bound 0.345	0.12 (0.03 to 0.48) **0.12 (0.03** to **0.54)**	0.69 (0.17 to 2.77) **0.69 (0.15** to **3.14)**	0.15	0.78	0.07	0
RE with $(R_{OR} - 1) \sim gamma(2.62, 0.721)$ and $\tau_{OR} = \log(R_{OR})/3.92$	0.12 (0.03 to 0.47) **0.12 (0.02** to **0.56)**	0.71 (0.17 to 2.97) **0.71 (0.14** to **3.69)**	0.01	0.85	0.14	0

Example 2: TA336[36]. Change from baseline in body weight (kg) at 24 wk: treatment effect on mean difference scale	**MD, median (95% CrI) Empa 10mg + Met + SU vs. placebo + Met + SU**	**MD, median (95% CrI) Empa 10mg + Met + SU vs. linagliptin + Met + SU**	P_L	P_M	P_H	P_{EH}
FE	−1.77 (−2.18 to −1.35)	−2.10 (−2.64 to −1.54)	0	0	0	0
RE with $\tau_{MD} \sim uniform[0, 5]$	−1.76 (−6.10 to 2.70) **−1.76 (−7.70 to 4.38)**	−2.08 (−8.12 to 4.08) **−2.08 (−10.75 to 6.55)**	0.14	0.32	0.19	0.35
RE with $\tau^2_{OR} \sim lognormal(0.256, 1.74^2)$ and $\tau_{MD} = 2.61 \times \tau_{OR}/1.81$	−1.77 (−2.88 to −0.63) **−1.77 (−3.27 to −0.18)**	−2.10 (−3.65 to −0.51) **−2.10 (−4.22 to 0.13)**	0.18	0.70	0.10	0.02
RE with $\tau^2_{OR} \sim lognormal(0.256, 1.74^2)$ truncated with upper bound 0.345 and $\tau_{MD} = 2.61 \times \tau_{OR}/1.81$	−1.77 (−2.62 to −0.93) **−1.77 (−2.95 to −0.63)**	−2.10 (−3.30 to −0.93) **−2.10 (−3.74 to−0.46)**	0.21	0.75	0.04	0
RE with $(R_{OR} - 1) \sim gamma(1.94, 0.741)$ and $\tau_{MD} = 2.61 \times \log(R_{OR})/(3.92 \times 1.81)$	−1.78 (−2.76 to −0.80) **−1.77 (−3.11 to −0.45)**	−2.10 (−3.47 to −0.72) **−2.10 (−3.98 to−0.23)**	0.08	0.88	0.03	0

FE, fixed effect; RE, random effects; OR, odds ratio; CrI, credible interval; MD, mean difference; Empa, empagliflozin; Met, metformin; SU, sulphonylurea.

[a] P_L, P_M, P_H and P_{EH} denote the probability that heterogeneity being low, moderate, high, and extremely high, respectively. Truncated log normal distribution has upper bound 0.345 representing that the "range" of odds ratios between studies cannot exceed 10. Results in bold are the predictive distributions of the effects of treatments in a new study.

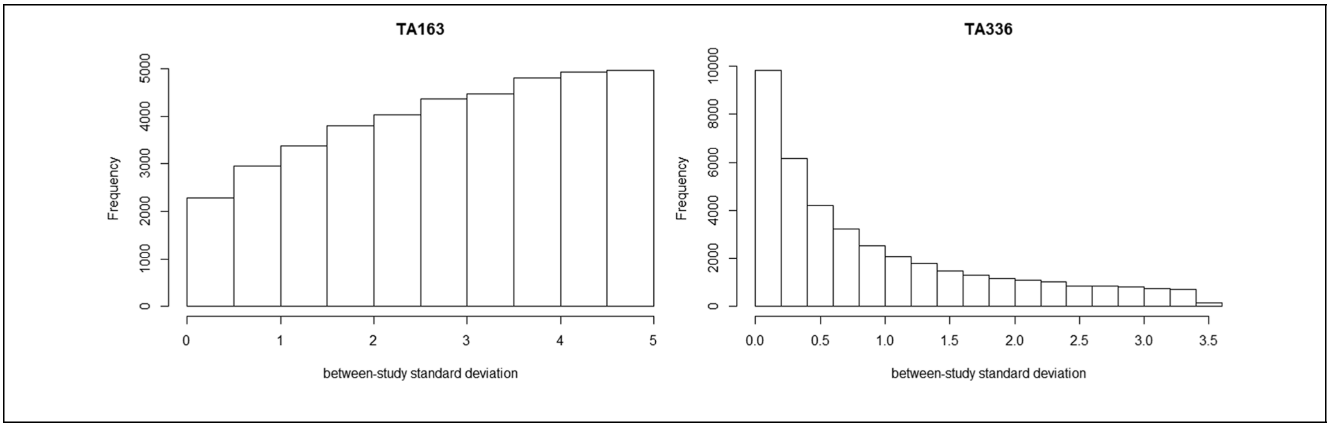

Figure 4 Posterior histogram plot of the between-study SD using prior distribution as uniform [0,5].

unlikely to be plausible, and the results using this prior distribution would not lead to reasonable posterior beliefs.

Results using empirical evidence as prior distribution and the elicited prior distributions for the heterogeneity were much less uncertain than those produced using the uniform prior distribution but, as expected, differed depending on which prior distribution was used (Table 2). Using the untruncated lognormal prior, there was a small posterior probability that heterogeneity was extremely high, 0.08. The truncated lognormal and elicited prior distributions for the heterogeneity parameter both provided zero posterior probability of extreme values for the between-study SD. The truncating eliminated the possibility of extreme heterogeneity; i.e., the largest OR in one study could be no more than 10 times the OR in another study. The elicited prior distribution can be found in Appendix 3, which resulted in the probability of heterogeneity being low, moderate, and high as 0.01, 0.85, and 0.14, respectively. The analyses using informative prior distributions for the heterogeneity parameter all suggested that ciclosporin reduced the colectomy rate at 3 mo compared with placebo based on both the credible and prediction intervals, but the effect of infliximab v. placebo was inclusive. The credible and predictive intervals in the analyses using empirical evidence and elicited prior distributions were wider than the fixed effect interval because of the extra uncertainty but more plausible than analyses based on the fixed effect model and random effects model with a uniform prior distribution.

TA336[36] was a technology appraisal of empagliflozin for the treatment of type 2 diabetes mellitus. Data were available from 6 studies of 8 treatments in combination with metformin (Met) or metformin and sulphonylurea (Met+SU) (placebo, 10 and 25 mg empagaliflozin, linagliptin, sitagliptin, saxagliptin, and 100 and 300 mg canagliflozin). The outcome measure re-analyzed here is the change from baseline in body weight for the third line treatment of type 2 diabetes mellitus at 24 wk. A fixed effect model was used in the original submission.

Table 2 presents the results of 10 mg empagliflozin + Met + SU v. placebo + Met + SU and linagliptin + Met + SU as an illustration. The DIC statistics for the 5 models were again similar: 3.82, 4.94, 4.61, 5.01, and 4.65. The fixed effect model showed that 10 mg empagliflozin +Met + SU reduced the change from baseline in body weight compared with placebo + Met + SU and linagliptin + Met + SU. When using the uniform [0, 5] prior distribution for the heterogeneity parameter, there was more "updating" in this case (Figure 4) but still a large posterior probability, 0.35, that the heterogeneity was extremely large. There was a small probability, 0.02, that the heterogeneity was extremely large when untruncated lognormal was used as the prior. The truncating eliminated the possibility of extreme heterogeneity in treatment effects between studies. The elicited prior can be found in Appendix 3, which resulted in the probability of heterogeneity being low, moderate and high as 0.06, 0.88, and 0.06, respectively. The analyses using informative prior distributions for the heterogeneity parameter all suggested that empagliflzin 10 mg is associated with a beneficial treatment effect as compared with placebo or linagliptin (all in combination with Met and SU) based on the credible and prediction intervals, except for the prediction interval using untruncated lognormal prior.

Discussion

Our review of NICE STAs showed that 17 (71%) of 24 fixed effect NMAs were chosen on the basis that there were too few studies with which to estimate the heterogeneity parameter, but not that there was unlikely to be heterogeneity or that a conditional inference was of interest. A consequence of this is that decision uncertainty may be underestimated. The choice between using a fixed effect or random effects MA model depends on the inferences required and not on the number of studies. Although a fixed effect model is informative in assessing whether treatments were effective in the observed studies, when we expect heterogeneity between studies and want to make unconditional inferences and predictions about the treatment effect in a new study, a random effects model should be used.

When heterogeneity is expected, the simple framework we have proposed overcomes the inappropriate assumption behind the use of a fixed effect model. We argue that, in the absence of sufficient sample data, a minimum requirement should be to exclude extreme and implausible values from the prior distribution and the common choice of the prior distribution, such as uniform [0, 5] or [0, 2], should not be used. We have shown in the examples that the use of a uniform prior distribution when data are sparse would result in an implausible estimate for the heterogeneity parameter and unreasonable results for the treatment effect.

Our proposed elicitation framework is flexible with the amount of information provided by an expert. The minimum information required from the expert is the maximum possible value of the "range" of treatment effects on the natural scale. For example, if the additive treatment effect is an LOR, then the expert is asked whether the OR in one study could be x times that of the OR in another and what the maximum plausible value of x could be. If the expert is not able to provide any judgments on the "range" of treatment effects, then empirical evidence, such as a prior distribution proposed for the heterogeneity expected in future MAs,[15–18] could be considered. When the expert provides only the maximum value of the "range" of treatment effects, the prior distributions proposed by other authors[15–18] should be truncated accordingly before use in the analysis. Note that the truncation of the prior prosed by Rhodes et al.[18] requires transformation between the standardized mean difference scale and the odds ratio scale. If the expert provides complete probability judgments, then our proposed framework could facilitate the elicitation exercise. In terms of presenting the results, we propose reporting the prior and posterior probabilities of heterogeneity being low, moderate, high and extremely high rather than simply as the point estimate and the credible interval, thereby presenting more information about the consequences of the chosen prior distribution. We also advocate the use of prediction intervals for the treatment effects as proposed by others.[2,3,8] Prediction intervals provide a summary of the treatment effect expected in a new study, which is more relevant to decision making.

In summary, in the absence of sufficient sample data, it is important to incorporate genuine prior information about the heterogeneity parameter in a random effects pairwise MA/NMA. Eliciting probability judgments from experts is not straightforward but is important if the aim is to genuinely represent uncertainty in a justifiable and transparent manner to properly inform decision making. Our proposed elicitation framework uses external information, such as empirical evidence and experts' beliefs, in which the minimum requirement from the expert is the maximum value of the "range" of treatment effects. The method is also applicable to all types of outcome measures for which a treatment effect can be constructed on an additive scale.

Supplementary Material

Supplementary material for this article is available on the *Medical Decision Making* Web site at http://journals.sagepub.com/home/mdm.

References

1. Borenstein M, Hedges L V., Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods*. 2010;1:97–111.
2. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ Br Med J*. 2011;342:d549.
3. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc*. 2009;172:137–159.
4. Rice K, Higgins JPT, Lumley T. A re-evaluation of fixed effect(s) meta-analysis. *J R Stat Soc Ser A*. 2017. doi:10.1111/rssa.12275.
5. Higgins J, Thompson S, Deeks J, Altman D. Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. *J Health Serv Res Policy*. 2002;7:51–61.
6. Dias S, Welton NJ, Sutton AJ, Ades AE. Evidence synthesis for decision making 1: Introduction. *Med Decis Making*. 2013;33:597–606.
7. Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2: A generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Med Decis Making*. 2013;33:607–617.
8. Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence synthesis for decision making 3: Heterogeneity–subgroups,

meta-regression, bias, and bias-adjustment. *Med Decis Making*. 2013;33:618–640.

9. Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials. *Med Decis Making*. 2013;33:641–656.
10. Dias S, Welton NJ, Sutton AJ, Ades AE. Evidence synthesis for decision making 5: The baseline natural history model. *Med Decis Making*. 2013;33:657–670.
11. Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence synthesis for decision making 6: embedding evidence synthesis in probabilistic cost-effectiveness analysis. *Med Decis Making*. 2013;33:671–678.
12. Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Anal*. 2006;1:515–533.
13. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *J R Stat Soc Ser B*. 2002;64:583–639.
14. Higgins JPT, Whitehead A. Borrowing strength from external trials in a meta-analysis. *Stat Med*. 1996;15:2733–2749.
15. Pullenayegum EM. An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes. *Stat Med*. 2011;30:3082–3094.
16. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JP. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *Int J Epidemiol*. 2012;41:818–827.
17. Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JPT. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Stat Med*. 2015;34:984–998.
18. Rhodes KM, Turner RM, Higgins JPT. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *J Clin Epidemiol*. 2015;68:52–60.
19. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: A comparative study. *Stat Med*. 1995;14:2685–2699.
20. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester: John Wiley & Sons, Ltd; 2004.
21. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol*. 1997;50:683–691.
22. NICE. Dapagliflozin in combination therapy for treating type 2 diabetes; 2013. [Accessed January 18, 2017]. Available from: https://www.nice.org.uk/guidance/ta288
23. NICE. Apixaban for the treatment and secondary prevention of deep vein thrombosis and/or pulmonary embolism; 2015. [Accessed January 18, 2017]. Available from: https://www.nice.org.uk/guidance/ta341
24. NICE. Tenofovir disoproxil for the treatment of chronic hepatitis B; 2009. [Accessed January 18, 2017]. Available from: https://www.nice.org.uk/guidance/ta173
25. NICE. Obinutuzumab in combination with chlorambucil for untreated chronic lymphocytic leukaemia; 2015. [Accessed January 18, 2017]. Available from: https://www.nice.org.uk/guidance/ta343
26. European Food Safety Authority. Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA J*. 2014;12:3734.
27. Morgan MG (Millett G, Henrion M, Small MJ. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge: Cambridge University Press; 1990.
28. O'Hagan A. *Uncertain Judgements: Eliciting Experts' Probabilities*. Chichester: Wiley; 2006. p 321.
29. Alhussain ZA, Oakley JE. *Eliciting Judgements about Uncertain Population Means and Variances*. 2017. Available from: https://arxiv.org/pdf/1702.00978.pdf.
30. Oakley JE. SHELF: Tools to Support the Sheffield Elicitation Framework. R package version 1.2.3. [Internet]. 2017. Available from: https://cran.r-project.org/package = SHELF
31. Gore SM. *Biostatistics and the Medical Research Council. Med Res Counc News*. 1987;35:19–20.
32. Friedrich JO, Adhikari NK, Beyene J, et al. The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: A simulation study. *BMC Med Res Methodol*. 2008;8:32.
33. Friedrich JO, Adhikari NKJ, Beyene J. Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. *J Clin Epidemiol*. 2011;64:556–64.
34. Lumley T. Network meta-analysis for indirect treatment comparisons. *Stat Med*. 2002;21:2313–2324.
35. NICE. Infliximab for acute exacerbations of ulcerative colitis [Internet]. NICE; [Accessed January 18, 2017]. Available from: https://www.nice.org.uk/guidance/ta163
36. NICE. Empagliflozin for the treatment of Type 2 diabetes mellitus (T2DM) [Internet]. Available from: https://www.nice.org.uk/guidance/TA336.
37. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Stat Comput*. 2000;10:325–337.