Journal of Clinical Epidemiology 109 (2019) 51–61

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Different evidence summaries have implications for contextualizing findings of meta-analysis of diagnostic tests

Anja Zgodic[a,*], Christopher H. Schmid[a,b], Ingram Olkin[c,†], Thomas A. Trikalinos[a,d]

[a]*Center for Evidence Synthesis in Health, School of Public Health, Brown University, Providence, RI, USA*
[b]*Department of Biostatistics, School of Public Health, Brown University, Providence, RI, USA*
[c]*Department of Statistics, Stanford University, Palo Alto, CA, USA*
[d]*Department of Health Services, Policy & Practice, School of Public Health, Brown University, Providence, RI, USA*

Accepted 8 January 2019; Published online 15 January 2019

Abstract

Objective: To evaluate diagnostic tests, analysts use meta-analyses to provide inputs to parameters in decision models. Choosing parameter estimands from meta-analyses requires understanding the meta-analytic and decision-making contexts.

Study Design and Setting: We expand on an analysis comparing positron emission tomography (PET), PET with computed tomography (PET/CT), and conventional workup (CW) in women with suspected recurrent breast cancer. We discuss Bayesian meta-analytic summaries (posterior mean over a set of existing studies, posterior estimate in an existing study, posterior predictive mean in a new study) used to estimate diagnostic test parameters (prevalence, sensitivity, specificity) needed to calculate quality-adjusted life years in a decision model contextualizing PET, PET/CT, and CW.

Results: The mean and predictive mean give similar estimates, but the latter displays greater uncertainty. Namely, PET/CT outperforms CW on average but may not do better than CW when implemented in future settings.

Conclusion: Selecting estimands for decision model parameters from meta-analyses requires understanding the relationship between decision settings and meta-analysis studies' settings, specifically whether the former resemble one or all study settings or represents new settings. We provide an algorithm recommending appropriate estimands as input parameters in decision models for diagnostic tests to obtain output parameters consistent with the decision context.

Keywords: Decision analysis; Prevalence; Sensitivity; Specificity; Quality-adjusted life years; Predictive mean

1. Introduction

Meta-analyses of medical tests often require advanced contextualization because such analyses usually summarize test performance (test accuracy), which is a surrogate for patient-relevant clinical outcomes. Indeed, most effects of testing are potentially indirect: test results can change diagnostic thinking and treatment decisions, which in turn may influence patient outcomes. To interpret a meta-analytic summary of test performance, it is often necessary to contextualize the findings by placing them in an appropriate frame of reference. This is typically done by using decision modeling to assess the impact of alternative testing strategies on clinical outcomes [1,2]. The health care community at large has recognized the importance of asking meaningful questions when planning these evidence reviews combined with decision analyses and relies on established stakeholder-driven processes to develop the key questions of its reviews. By contrast, the nuances of contextualizing results, that is, describing the implications of an evidence synthesis as they apply to a particular setting, have received little attention.

The example in Box 1, based on a technology assessment commissioned by the UK's National Institute for Health Research (NIHR) [3], motivates the methodological question of exactly which meta-analytic estimand to use to inform decision- or policy-making. The assessment includes a meta-analysis of the performance of three tests in diagnosing recurrent breast cancer (positron emission

This work was supported in part by the Evidence-based Practice Centers (EPC) Program of the Agency for Healthcare Research and Quality (Contract Number to the Brown University EPC HHSA290201500002I – Task Order 1; Principal Investigator TA Trikalinos).

Conflict of interest: The authors have no conflicts of interest to report.

† Deceased.

* Corresponding author. Center for Evidence Synthesis in Health, School of Public Health, Brown University, 121 S Main Street, Providence, RI, 02912, USA. Tel.: +1 4015858350; fax: +1 4018632900.

E-mail address: anja_zgodic@alumni.brown.edu (A. Zgodic).

https://doi.org/10.1016/j.jclinepi.2019.01.002
0895-4356/

What is new?

Key findings

- Propagation of outcomes from decision models that use parameter estimates derived from meta-analyses can vary considerably based on the parameter estimand used.

What this adds to what is known?

- We offer an algorithm to assist in determining which meta-analysis estimands to use as input parameters in decision models for diagnostic tests to ensure accurate output parameters that are consistent with the decision context and allow for an accurate evaluation of said diagnostic tests.

What is the implication and what should change now?

- Analysts must consider whether the setting in which the decision is to be applied is the same as or similar to that of an existing study in the meta-analysis or whether it is a new setting that may or may not be informed by the meta-analysis.

tomography [PET] alone, PET in combination with computed tomography [PET/CT], or conventional workup [a diverse set of other diagnostic procedures]) that synthesized studies in which the true disease status was assessed by histology or long-term clinical follow-up [3]. As shown in Box 1, the point estimates for the meta-analytic means suggest that PET/CT is more sensitive but less specific than PET and that both PET/CT and PET are more sensitive and more specific than conventional workup in diagnosing recurrence. However, it is unclear what the tests' diagnostic performance implies in terms of expected clinical outcomes with different testing modalities, a critical input to decision-making about which test to use. At a more fundamental level, it is unclear *which summary* from a meta-analysis should be used to estimate the expected clinical outcomes and thus inform decision-making.

The first section of this report discusses alternative summaries one can generally obtain from a meta-analysis, their meaning, and how to choose among them [5,6]. We then describe a simple decision modeling exercise, inspired by a decision model from Auguste et al.'s (2011) [4] work, that contextualizes meta-analytic evidence summaries based on our extension of Pennant et al.'s (2010) [3] NIHR technology assessment. This decision model incorporates many simplifying assumptions that collapse meta-analytic results for true positive rate (TPR), false positive rate (FPR), and disease probability into an all-encompassing measure called quality-adjusted life years (QALYs). We use this decision model and its QALYs metric exclusively as a device to illustrate our methodological points, although doing so reduces the clinical relevance of results.

In our extension of the work by Pennant et al. (2010) [3], we assume that the systematic review of medical test performance has been conducted following standard methodological guidance [1,2], and we apply minimal decision modeling (using the meta-analysis estimates) to compare the expected clinical utility of various tests in order to interpret the results of their meta-analysis. A practical question is then which estimate of TPR (i.e., sensitivity), FPR (i.e., one minus specificity), and disease probability (prevalence) to use in the decision model to contextualize the interpretation. For example, policy-makers from a US institution may wish to use a meta-analysis to inform their decision on whether to cover the patient costs for a particular medical technology. Presumably, they would be interested in the findings of the most applicable studies (e.g., recent, large, well-conducted US-based studies in older adults), but would also want to take into account the remainder of the evidence base. At one extreme, they could consider only the most applicable study, ignoring other studies if these do not provide useful information. Or, they could use the average of the subgroup of studies conducted, say in the United States, which might be preferable to the average of all studies, if some are foreign and therefore reflect other environments. Using a (subgroup) average in a predictive sense also ignores the uncertainty introduced by a new (future, or different setting/population) implementation of a medical test.

Generally, the choice of meta-analytic estimand has implications for the interpretation of the conclusion, and on the certainty with which conclusions can be drawn [6]. We elaborate on these implications and their certainty throughout our report. We opted to expand on the example by Pennant et al. (2010) [3] and Auguste et al. (2011) [4] because it is fully quantitative and therefore allows precise expositions of the relevant issues. However, our insights can be fully extended to systematic reviews that do not involve test performance and systematic reviews without meta-analysis.

2. Methods

Throughout this text, we assume that meta-analyses are based on systematic reviews of the literature. For economy of language, we will occasionally loosely refer to a "meta-analysis" as a shorthand for a "meta-analysis of studies identified in a systematic review." We will also refer to "eligibility criteria of a meta-analysis" but mean "eligibility criteria for the question addressed in the systematic review via a meta-analysis of (a subset of) the included studies."

2.1. Meta-analysis models

Studies included in a meta-analysis are almost always clinically, epidemiologically, and methodologically diverse.

Box 1 PET/CT, PET, and conventional workup for the diagnosis of recurrent breast cancer.

After the initial treatment of breast cancer, women are typically followed with clinical examinations and mammography to detect local recurrence of treatable disease for at least 5 years. Women with a history of breast cancer who develop symptoms should be further diagnosed, to establish whether the symptoms are innocent, secondary to local recurrence, which is treatable with curative intent, or due to terminal metastatic disease, in which the goal of treatment is to improve the length and quality of life. These women may undergo conventional workup (CW) or can be tested with whole body positron emission tomography (PET) or its combination with computed tomography (PET/CT). Depending on the presenting symptoms, the CW comprises widely available and accessible investigations such as X-rays, CT, or magnetic resonance imaging, for checking a range of tissues; ultrasound, to check for liver metastases; and bone scintigraphy, to check for metastases in the bones.

Over the last 2 decades, whole body PET/CT or PET has been increasingly used in the diagnosis and management of cancers. Compared to CW, PET/CT and PET are substantially more expensive and much less available. PET/CT may be more sensitive but less specific than PET in detecting both local and metastatic diseases. Both PET/CT and PET may have better diagnostic accuracy than CW. Because accurate detection of breast cancer recurrence may be associated with less advanced and better treatable disease, the three diagnostic strategies can effect different clinical outcomes. Furthermore, understanding the tradeoffs between the expected benefits (longer and better life) and costs associated with infrastructure development for PET with or without concurrent CT is an important input for health care decision-making. In this report, we focus on comparing the expected benefits with the three diagnostic strategies.

Pennant et al. (2010) [3] conducted a meta-analysis of diagnostic test accuracy for the three tests synthesizing studies where the reference standard used to define the true disease status was histological diagnosis or long-term clinical follow-up. We use their data to perform meta-analyses of test performance in a Bayesian setting. Based on the meta-analysis, the random effects mean sensitivity and specificity are as follows:

- for PET/CT (data from five studies), sensitivity = 0.950 (95% credible interval: 0.876 to 0.986), specificity = 0.867 (0.736 to 0.952), correlation of sensitivity, and specificity between studies = 0.070 (−0.942 to 0.954)
- for PET (data from 13 studies), sensitivity = 0.873 (0.823 to 0.912), specificity = 0.906 (0.829 to 0.961), and correlation = −0.470 (−0.980 to 0.663), and
- for conventional workup (data from 11 studies), sensitivity = 0.789 (0.717 to 0.854), specificity = 0.800 (0.633 to 0.921), correlation = −0.200 (−0.967 to 0.774).

We then construct a decision analysis based on Auguste et al. (2011) [4] and use results from our meta-analyses of diagnostic accuracy as estimates of the sensitivity and specificity of testing options in the decision model.

Our focus is methodological rather than clinical. Which quantities from the meta-analysis should we use to inform our decision analysis? We argue against the common practice of using the random effects meta-analysis means to inform the sensitivity and specificity in the decision model. We explain why different statistics from a meta-analysis should be used and provide guidance depending on the context of the analysis.

To allow for this unexplained between-study heterogeneity, it is common to use random effects meta-analysis models in evidence synthesis. Empirical evidence shows that such heterogeneity is almost always present with meta-analyses of diagnostic test performance [7]. A Bayesian framework is a natural tool for accounting for this heterogeneity and we use it, as well as its accompanying posterior estimates from random effects meta-analyses, in this report. Nonetheless, these models may also be fit and conclusions drawn for contextualization in a non-Bayesian framework with corresponding parameter estimates.

A typical random effects meta-analysis partitions the variance in the data into two parts. Within-study variance describes the uncertainty associated with estimates of key parameters in each study. Between-study variance describes the inherent variability in the study parameters across studies. Within-study variance describes the sampling variance of the study estimate and therefore decreases as more data are collected in each study. This in turn increases the precision of study parameter estimates. Between-study variance is a characteristic of inherent heterogeneity. It is more precisely estimated with more studies. A common model assumes a normal distribution for both types of variation, that is,

$$y_i \sim N\left(\delta_i, \sigma_i^2\right) \text{ and } \delta_i \sim N\left(\Delta, \tau^2\right),$$

where y_i is the observed study-level parameter (e.g., log-odds ratio, mean difference, or transformed proportion, such as logit sensitivity or specificity) in study i and σ_i^2 is the corresponding sampling variance; δ_i is the unobserved study-specific true parameter, and Δ and τ^2 denote the mean and variance of the distribution of true effects. Usually, σ_i^2 is assumed known. The estimate of Δ along with

a confidence or credible interval is the typically reported meta-analysis summary. Our best prediction about the expected true effect in a future setting is the predictive distribution $\delta_{new} \sim N(\Delta, \tau^2)$. The predictive confidence or credible interval for δ_{new} is wider than that for Δ because it incorporates the random effects variance τ^2.

We illustrate these concepts using the Pennant et al. (2010) [3] study that examined the performance of PET, PET/CT, and conventional workup (CW) technologies to aid the NIHR's decision-making process with regard to implementing those technologies for diagnosing recurrent breast cancer in the UK. We supplement the original study by performing an additional meta-analysis on prevalence. Both the prevalence as well as the test performance meta-analysis are based on 15 diagnostic cohort studies (note: the systematic review by Pennant et al. (2010) [3] excluded diagnostic case-control studies), where five examine PET/CT, 13 examine PET, and 11 examine CW (Table 1) [3]. Most of the studies examine two or more tests applied to the same patients, but we treat them as if they applied each test to a different set of patients. While this simplification results in some loss of information, it yields point estimates of average test performance close to those from an analysis that models the cross-classification of test results [8]. In case-control studies, prevalence is fixed and only test performance needs to be modeled; in cohort studies, prevalence must also be modeled. Some approaches model test performance and prevalence jointly [9]. Here, we replicate

Table 1. Data on the cross-classification of test results and breast cancer recurrence status

			Recurrence, D		No recurrence, $\overline{D}$	
ID	**Author (y)**	**Test name**	**Test (+)**	**Test (−)**	**Test (+)**	**Test (−)**
1	Dirisamer (2010)	PET/CT	40	2	0	10
2	Haug (2007)	PET/CT	24	1	1	8
3	Radan (2006)	PET/CT	17	3	4	13
4	Veit-Haibach (2007)	PET/CT	19	0	4	21
5	Fueger (2005)	PET/CT	31	2	4	21
6	Bender (1997)	PET	10	4	2	48
7	Goerres (2003)	PET	11	3	1	17
8	Hathaway (1999)	PET	6	0	0	1
9	Dirisamer (2010)	PET	34	8	0	10
10	Fueger (2005)	PET	28	5	7	18
11	Haug (2007)	PET	23	3	1	7
12	Veit-Haibach (2007)	PET	17	2	6	19
13	Abe (2005)	PET	14	0	1	29
14	Ohta (2001)	PET	7	2	1	42
15	Hubner (2000)	PET	36	6	6	16
16	Gallowitsch (2003)	PET	33	1	5	23
17	Vranjesevic (2002)	PET	39	3	3	16
18	Wolfort (2006)	PET	13	3	0	7
19	Abe (2005)	CW	11	3	0	30
20	Ohta (2001)	CW	7	2	8	34
21	Raileany (2004)	CW	7	0	3	10
22	Gallowitsch (2003)	CW	28	6	13	15
23	Vranjesevic (2002)	CW	33	9	6	13
24	Wolfort (2006)	CW	11	4	0	7
25	Dirisamer (2010)	CW	28	14	0	10
26	Haug (2007)	CW	24	2	2	6
27	Radan (2006)	CW	14	6	9	9
28	Veit-Haibach (2007)	CW	17	2	6	19
29	Hubner (2000)	CW	22	9	6	7

Abbreviations: CT, computed tomography; CW, conventional workup; PET, positron emission tomography.

Data were obtained from Pennant et al. (2010) [3]. In studies where data were available for more than one test, where possible, Pennant et al. selected only data for participants who had undergone all such tests. Nonetheless, we see three slight exceptions to this: Haug (2007), Hubner (2000), and Wolfort (2006). No attempts have been made to verify the completeness or correctness of these data.

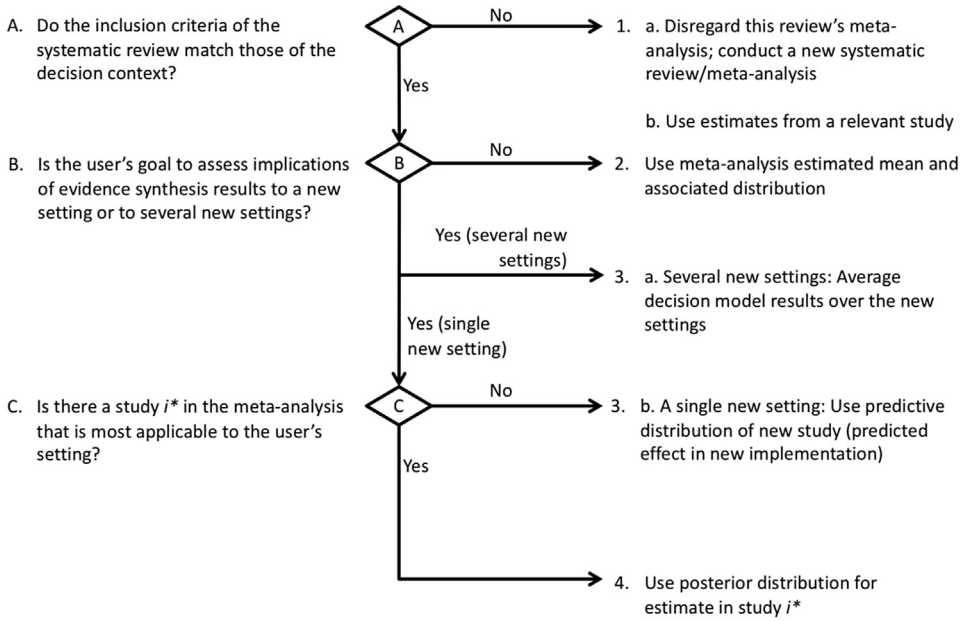

Fig. 1. Choice of estimates from an evidence synthesis depending on users' purpose. The algorithm pertains to any evidence synthesis that will be used to inform decision-making. This includes scenario analyses and conditional sensitivity analyses that examine the impact of alternative modeling assumptions and the influence of model inputs on outputs, respectively. In this article, it applies to estimates of prevalence or of sensitivity and specificity.

Pennant et al.'s (2010) [3] model for test performance and construct a separate model for prevalence showing how prevalence and test performance can be used together to obtain the predicted effect in a new implementation of these diagnostic tests. These models are outlined in full detail in Appendix. Overall, the meta-analysis studies include data on 604 true positive (TP), 105 false negative (FN), 99 false positive (FP), and 486 true negative (TN) test results.

2.2. Meta-analysis summaries

On completing a new or using an existing meta-analysis as delineated in Appendix, decision analysts must contextualize the resulting estimates and their differences. Fig. 1 presents a structured way to proceed based on proper statistical interpretation of the various quantities estimated in a meta-analysis [5,6,10–12]. The analyst must first determine whether the meta-analysis is relevant to the decision informing the implementation of a diagnostic test (question A). Sometimes, the inclusion criteria used to select the studies in a systematic review, and thus in a meta-analysis, do not match those defining the context of the decision analysis. This is most common when a decision analysis obtains parameter estimates from previous systematic reviews carried out for other reasons, or when it considers scenario (sensitivity) analyses in diverse decisional contexts. In such cases, the overall meta-analysis will not be useful to the decision, although one or more of its studies may be [13]. For example, regulators in a specific country may need to make a decision and may wish to only use results of studies conducted in that country. Or, the regulators may wish to focus on use of a treatment for a particular indication when the meta-analysis addresses use of the treatment for a wider class of indications. The decision analyst may then choose to conduct a new meta-analysis of a (sub)set of studies more relevant to the decision or ignore all the studies in the meta-analysis and search for another way to obtain the information.

Option 1 in Fig. 1 represents the case where question A is answered in the negative; the decision analyst does not perceive the systematic review and associated meta-analysis as useful and decides to disregard it. An additional case arises in this scenario: one particular study may be applicable and relevant to the decision context. In that case, using only that specific study estimate to inform the decision-making process can be a viable alternative and is represented by option 1b in Fig. 1 [6]. We do not discuss the options when the meta-analysis is not deemed useful (options 1a and 1b) further but instead focus on situations where the meta-analysis is relevant to the decision problem.

If the decision analyst makes the decision to use the meta-analysis, question B requires assessing whether interest lies only in a summary of the evidence, or in informing a decision about a new implementation of the test or intervention at hand. If only a summary is needed, the random effects meta-analysis mean Δ and the associated random effects distribution suffice (option 2). A "new implementation" may be a new setting (e.g., implementing PET/CT, PET, or CW-based patient management protocols at Rhode Island Hospital) or a set of new settings (e.g., implementing the same testing protocols across the National Health System in the UK, which comprises many hospitals).

We start with the case where the user is interested in generalizing to a single new setting. Then, question C defines the new setting to the extent feasible. Opportunities to do so are typically limited in literature-based meta-analyses because one is working with summary data. (The meta-analysis model described previously has no covariates and so meta-regression is not considered further; however, an analogous rationale can be followed when meta-regression is possible.) At best, one can identify a particular study, say study i^*, which is most applicable to the setting of interest. In other words, the clinical setting, the distribution of population characteristics, the testing protocols (in terms of the exact version of the tests, the positivity cutoffs, the training of the readers), and so on, in study i^* are observationally equivalent to the target setting. The posterior estimate of the true effect δ_{i*} in study i^* would then be the best prediction for a new implementation in the setting of interest, assuming that study i^* is a priori exchangeable with the other studies and that the meta-analysis model is (approximately) correct (i.e., question C is answered in the affirmative, leading to option 4). The posterior estimate δ_{i*} is a weighted average of the observed effect y_{i*} in study i^* and the meta-analysis mean Δ. The weights depend on the relative precision with which y_{i*} and Δ are estimated with the more precisely estimated quantity receiving the greater weight. Sometimes, users are faced with a scenario where a subset of studies i^* is most applicable to the user's new setting, as opposed to one single study. In that case, the decision analyst would use the posterior distribution of the estimate in subset i^*.

If such a study cannot be identified (i.e., the answer to question C is negative, leading to option 3b), then a new (target) setting can be thought of as similar to the settings in all the studies observed (i.e., the distribution of effects in those studies accurately represents the likely value of the new effect). In other words, the user is interested in a predicted effect that is exchangeable with the effects in the studies included in the meta-analysis: this is the predicted true effect in a new study, δ_{new}. The uncertainty around δ_{new} incorporates between-study heterogeneity and is substantially larger than the uncertainty around the meta-analysis mean.

If implementation to a set of new settings is desired, the task is more involved. The decision model results

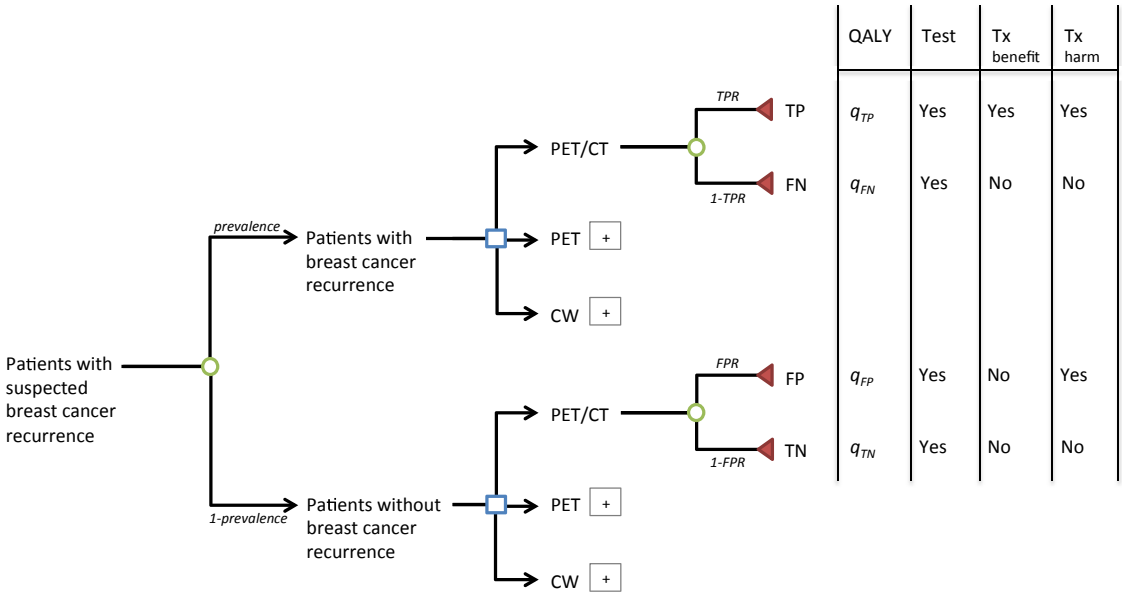

Fig. 2. Schematic of the decision tree used for contextualizing the results of the diagnostic tests meta-analysis. The user starts at the first node, where patients with suspected breast cancer recurrence separate into two groups, those with recurrence (with probability ''prevalence'') and those without recurrence (with probability ''1-prevalence''). Then, a decision is made on which diagnostic test to use to further assess the presence or absence of recurrence. Following the administration of the diagnostic test, the probability of receiving a TP, FN, FP, or TN test result depends on the TPR and FPR of the test at hand. The key quantities in the tree are the probability of recurrence (prevalence), which corresponds to the setting in which the tests would be used, and the TPR and FPR of the tests (test performance). The latter depend on the (latent) threshold for calling a test result positive vs. negative. In practice, even for qualitative tests such as the ones examined here, the threshold can be ''titrated'' after some period of field application of the tests. The columns on the right of the figure denote the QALY contributions (''QALY'') by test result: whether a test is used (''Test'') and whether patients are exposed to potential benefits (''Tx benefit'') or harms (''Tx harm'') from test-directed treatment. Note that this decision tree presents only potential benefits or harms for patients who engage in treatment following a positive test result. Potential benefits or harms for patients who do not engage in treatment are not considered in the context of this specific illustration. One could however easily extend the example to cover these types of benefits and harms as well. CT, computed tomography; CW, conventional workup; PET, positron emission tomography; FN, false negative result; FP, false positive result; TP, true positive result; TN, true negative result; FPR, false positive rate; TPR, true positive rate; QALY, quality-adjusted life year; Tx, treatment.

should be averaged over the predictions of the true effects in the new settings. Assuming that the distribution of random effects in the meta-analysis model describes the distribution of the true effects across all settings, it would suffice to (numerically) integrate the decision model results over the random effects distribution for the true effects of interest in the studies corresponding to the desired set of new settings (option 3a). We do not show computations for option 3a in the worked example because the integration over new settings is done in the decision analysis model (or equivalently, the decision analysis part of a ''comprehensive evidence synthesis'' model [14]). For this article, it is a computational complication that is not germane to the interpretation of meta-analysis results. For a worked example, see Welton et al. (2015) [6].

3. Contextualization example: diagnosis of breast cancer recurrence

3.1. Contextualization of meta-analysis estimates with a simple decision tree

As demonstrated in Appendix, options 2 through 4 from Fig. 1 lead to different point estimates and confidence/credible intervals and therefore to different inferences, conclusions, and clinical decisions. For further illustration, we continue our inquiry into the convenient example of diagnosis of breast cancer recurrence using PET alone, PET/CT, or CW. We guide the reader through a simple decision model for contextualizing meta-analysis results and demonstrate their differences with options 2 through 4 from Fig. 1. These results are obtained from the test performance and disease probability meta-analyses outlined in Appendix, using Bayesian methods with noninformative priors for the parameters of interest, and are presented here as posterior medians and central 95% credibility intervals.

The decision model consists of a simple decision tree modified from one described in the technology assessment conducted by Auguste et al. (2011) [4]. Fig. 2 shows the tree structure. For each of the two groups determined by the unknown disease recurrence status, a square decision node divides patients into three subgroups according to three test-and-treat strategies. The figure expands the strategy that uses PET/CT; the tree structure for the other strategies is the same. In each strategy, women in whom cancer may recur receive the test, which returns either a positive or negative result. These lead to true- and false-positive and -negative results depending on the eventual state of recurrence. Assuming that the test result determines whether further treatment is given (withhold when test is negative, administer when test is positive), the implications of actions induced by test results as well as associated clinical events then differ for true positive, false negative, false positive, and true negative test results and can be measured as QALYs. This metric measures time remaining in life, adjusted for its quality, and is commonly used to evaluate the utility of a medical intervention or treatment [15].

By starting at the terminal nodes of the decision tree–based model and working backward to the initial node multiplying appropriate quantities, we obtain expectations for effectiveness measures for each testing strategy. In principle, we would favor the test with the highest expected effectiveness, namely the highest expected QALYs.

In most applications, one would estimate prevalence from locally available data. However, here, to illustrate differences between the algorithms, we will get an estimate of prevalence from the systematic review. Let the estimated prevalence of recurrence be π' and the estimates of TPR and FPR in test j be TPR'_j and FPR'_j, respectively, as determined by whichever of options 2, 3b, or 4 (laid out in Fig. 1) is chosen. Specifically, π' might be equal to the meta-analysis mean for prevalence (π, option 2), the predictive distribution of prevalence in a new cohort (π_{new}, option 3b), and the posterior estimates of each study's prevalence (π_i, option 4). (Refer to Appendix for an in-depth elaboration on π, π_{new}, and π_i) TPR'_j and FPR'_j encompass the meta-analysis mean for the rates (TPR_j and FPR_j, option 2), the predictive distribution of prevalence in a new cohort ($TPR_{new,\ j}$ and $FPR_{new,\ j}$, option 3b), and the posterior estimates of each study's rates (TPR_{ij} and FPR_{ij}, option 4). (Refer to Appendix for an in-depth elaboration on TPR_j, FPR_j, $TPR_{new,\ j}$, $FPR_{new,\ j}$, TPR_{ij}, and FPR_{ij}.) Then, in a new fixed cohort of size N_{new} women, the expected number of TP, $E(c^{TP}_{new,j})$, or FP, $E(c^{FP}_{new,j})$, test results are given by

$$E\left(c^{TP}_{new,j}\right) = N_{new}\pi' TPR'_j, \text{ and} \tag{12}$$

$$E\left(c^{FP}_{new,j}\right) = N_{new}(1-\pi')FPR'_j. \tag{13}$$

To make ranking of testing strategies possible, one can use expected QALYs as an integrative measure of effectiveness. QALYs are meant to incorporate in a single measure all downstream effects of a test's results. Denote the QALYs for women with TP, FN, FP, and TN test results as q_{TP}, q_{FN}, q_{FP}, and q_{TN}, respectively. Then, the expected QALYs for a single patient receiving test j are:

$$\begin{aligned} Q_j = {} & \pi' TPR'_j q_{TP} + \pi'\left(1 - TPR'_j\right)q_{FN} \\ & + (1-\pi')FPR'_j q_{FP} + (1-\pi')\left(1 - FPR'_j\right)q_{TN}. \end{aligned} \tag{1}$$

For our example, we take q_{TP}, q_{FN}, q_{FP}, and q_{TN} to be fixed at the values used in the decision analysis by Auguste et al. (2011) [4]: 0.8737, 0, 13.8724, and 13.8724 QALYs per woman, respectively. These choices assume that the extra workup and anxiety due to FP results does not affect a patient's QALYs.

We present disease probability and test performance model outputs as well as their impact on QALYs according

to options 2 through 4 as in Fig. 1: estimate for the meta-analysis mean effect (option 2), predictive estimate for a new study (option 3b), or the estimate of the most applicable study (option 4). We conduct Markov chain Monte Carlo simulations to obtain posterior distributions for estimates from options 2, 3b, and 4, using data from studies presented in Table 1. We present the medians and central 95% credibility intervals of these posterior distributions, conditional on Table 1 data, as results.

3.2. Impact of alternative estimates on calculations about the effectiveness of testing

Fig. 3 shows the impact of meta-analytic options 2, 3b, and 4 on expected QALYs, sensitivity, specificity, and prevalence generated by the decision and meta-analysis models. For each measure, results based on the predictive estimate of future studies (white diamonds; option 3b) were substantially more uncertain compared to those based on the meta-analytic mean effect (black diamonds; option 2) for all three tests. The posterior distribution for each study (white circles; option 4), although differing from one study to another, also displayed more variation compared to the meta-analytic mean effect, but less variation compared to the predictive estimate. In essence, for each test, expected QALYs results using options 2, 3b, and 4 reflect results observed in QALYs components (sensitivity, specificity, and prevalence) when using options 2, 3b, and 4.

Using the meta-analytic mean and predictive estimate for the prevalence and TPR and FPR, Fig. 4 compares PET and CW against PET/CT for differences in total expected QALYs in a woman undergoing testing. Again, the predictive estimate of a future study (white diamonds; option 3b) resulted in substantially greater uncertainty

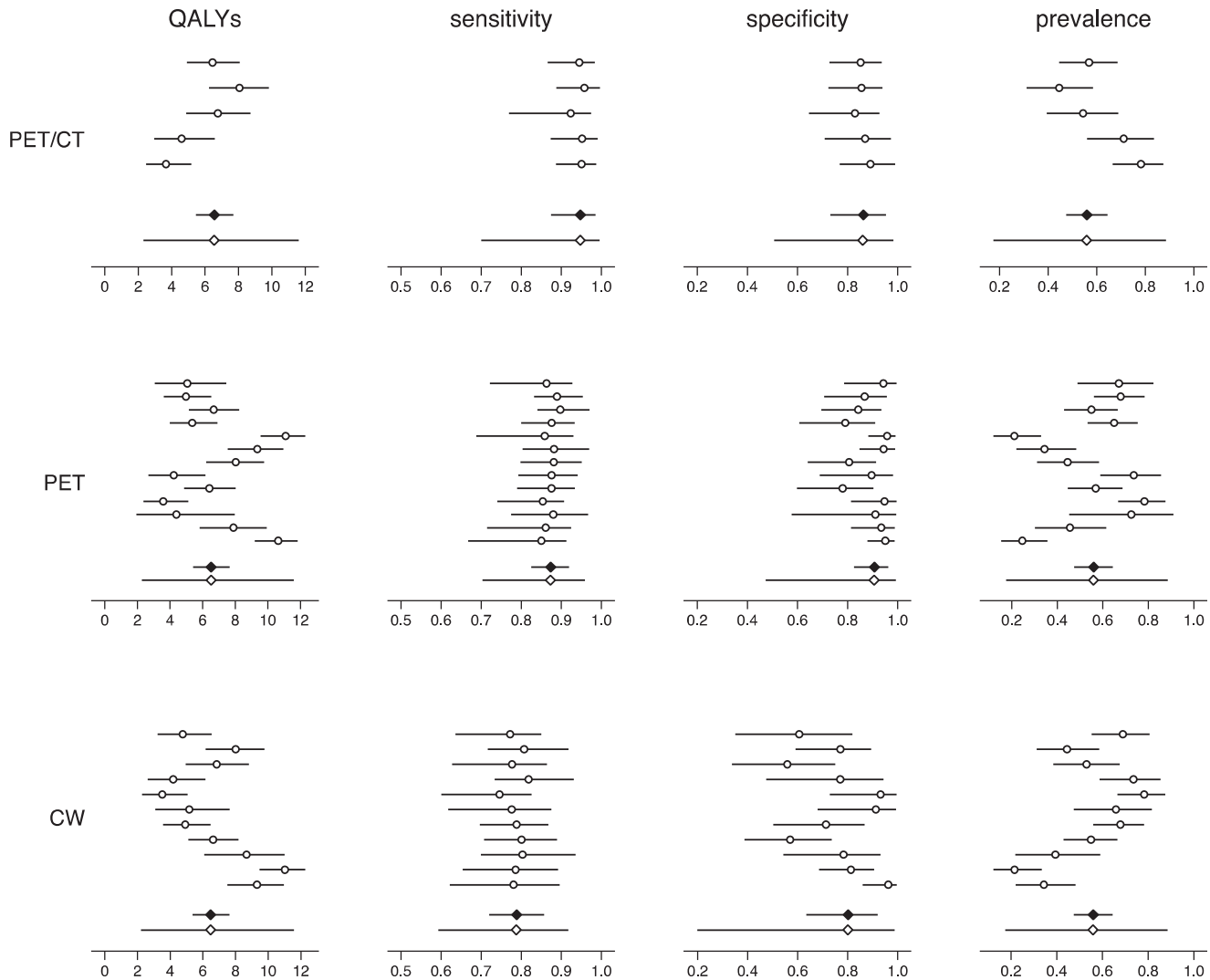

Fig. 3. Model results for total QALYs, sensitivity, specificity, and prevalence. Black diamonds: analyses using meta-analytic means (option 2). White diamonds: analyses using predictive estimates for new studies (option 3b). White circles: analyses using study-specific posterior estimates (option 4). Option 2 estimates are the same across the three prevalence plots because the prevalence meta-analysis was not test-specific; the same is true for prevalence option 3b estimates. There is one meta-analytic mean and one predictive estimate of the meta-analytic mean for prevalence, independent of the diagnostic test. CT, computed tomography; CW, conventional workup; PET, positron emission tomography.

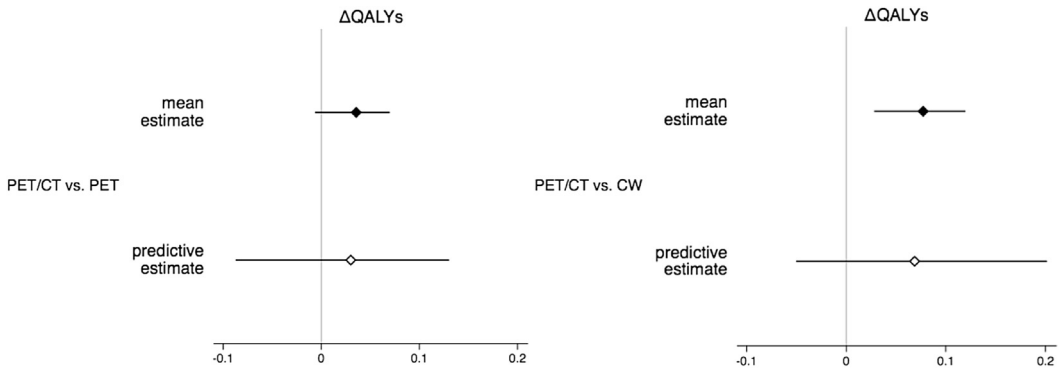

Fig. 4. Comparative test effectiveness from decision model. Black diamonds: analyses using meta-analytic means (option 2). White diamonds: analyses using predictive estimates for new studies (option 3b). Gray vertical lines denote no difference. Values greater than zero indicate improved outcomes with PET/CT. CT, computed tomography; CW, conventional workup; PET, positron emission tomography; $\Delta QALYs$, difference in total expected QALYs.

compared to results based on the meta-analytic mean (black diamonds; option 2). Using the mean estimate, one could state that, for instance, PET/CT outperforms CW. This conclusion only really applies, however, to the average setting and not to a new implementation, where the predictive credible interval encompasses zero. When comparing PET/CT with PET in the average setting and in a future setting, both credible intervals encompass zero, with the credible interval for a new implementation being considerably wider. As a result, one can draw the conclusion that PET/CT may not outperform PET in either current or future implementations for diagnosing breast cancer recurrence (but the likelihood is that it will).

4. Discussion

Systematic reviews and meta-analyses can be used to address a range of user needs. The most appropriate summary of the evidence depends on the user's intended purpose. For example, the random effects meta-analysis mean is useful for describing the evidence. However, when it comes to decisions about implementing a treatment or a test-based management strategy in a new setting, a predictive estimate (from a generic setting or in settings similar to those of specific included studies) may be more pertinent.

To date, most studies have compared the impact of different estimators of the meta-analytic mean on subsequent decision (or cost effectiveness) models [16,17]; few have assessed the use of other parameters of the model. Some studies, focusing on meta-analyses of treatment effectiveness, have highlighted the inappropriateness of the meta-analytic mean for guiding clinical decision-making or the design of future studies [5,10,11,18]. Welton et al. (2015) [6] and Ades et al. (2005) [10] considered the interpretation of decision models when treatment effects and baseline disease parameters are obtained by meta-analysis. Ades et al. (2005) [10] noted that "the mean treatment effect from a random effects meta-analysis will only seldom be an appropriate representation of the efficacy expected in a future implementation." Our work extends these observations to the meta-analysis and decision modeling of diagnostic test studies.

We chose a fully quantitative example on diagnostic tests for breast cancer recurrence because it allows us to be precise about the estimation and interpretation of the parameters of the evidence synthesis model used in the decision model. We have clearly distinguished the random effects meta-analysis mean Δ, the predictive estimate in a new setting/future implementation δ_{new}, and the prediction in a setting that is most similar to subset of studies i^* (δ_{i*}). In applied work, the random effects meta-analysis mean is often misinterpreted as the predicted effect in a new setting, although it ignores the uncertainty of a new implementation of the intervention or test at hand.

We would argue that the same nuances exist in a qualitative synthesis and decision-making process. However, qualitative processes are inherently less specific in their definitions (e.g., one would not distinguish between the three summaries of the evidence mentioned previously) and are much less conducive to describing the extent of uncertainty in the results of the evidence synthesis. Thus, although the same concepts apply, they are less transparent and can be all too easily ignored mistakenly. This is another manifestation of the information one loses when a quantitative analysis is not possible or indicated.

Some modeling is probably necessary in the assessment of tests because the value of testing is ultimately judged by its downstream effects [1,2]. Notwithstanding some, generally limited, direct effects of testing [19], in itself a positive or negative test result may indirectly affect diagnostic thinking and treatment decision-making, and, in turn, affect economic and health outcomes. Because studies of the downstream outcomes of test-and-treat strategies are rare, a tiered evaluation of medical tests is favored in technology assessments [20,21].

More practically, we offer the following generic suggestions for meta-analysts contemplating the use of decision analytic models to contextualize meta-analysis results:

- When there is extensive between-study diversity that is not successfully modeled through a meta-regression, *a meta-analysis−based estimate is arguably not appropriate, and one should only consider the results of the study that is most applicable to the decisional context.*
- *Consider meta-analysis−based estimates only for the subset of the studies in a systematic review*
  - (i) *that is most relevant to applied decisional contexts*. This could be all studies in the systematic review or the studies that use an error-free reference standard, implement a specific version of the test, enroll specific population strata, and so on.
  - (ii) *that successfully address the multitude of challenges that affect diagnostic test studies*, including disease spectrum effects [22], verification bias [23], reference standard with non-negligible misclassification error [24], and so forth. For a review of sources of bias in diagnostic test studies, see Whiting et al. (2004) [25]. An empirical manifestation of some of these challenges is that they induce between-study correlations between estimates of test performance metrics and prevalence [26]. The analysts should consider the impact that such issues have on the estimates of test performance, and, indirectly, on the clinical utility of testing. Ideally, bias-free or bias-corrected estimates should be used in decision models. For a discussion, see Trikalinos et al. (2016) [13].

- Depending on the scope of the decision modeling approach, *the predictive distribution of effects in future studies and meta-analysis−based study-specific posteriors are often appropriate inputs for populating decision models* [5,6,10,27,28]. In most cases, the meta-analytic mean effect is not the most appropriate input.
- *When a single study can be identified as particularly relevant to the target population of the systematic review, the meta-analysis−based study-specific posterior distribution can be used to populate the decision model* [6,10]. Model results obtained this way represent a compromise between two extremes: relying exclusively on a single study (while ignoring all other evidence) and using the meta-analytic mean (which may not be as applicable to the target population).
- *When investigators are interested in predicting the effects of future studies (or determining the impact of implementing a test in a new context), the predictive distribution of future studies should be used as an input for decision modeling* [4,10].
- *When investigators are interested in determining the impact of implementing a test in a set of new contexts, the decision model results should be averaged over the predictions of the true effects in the new settings* [6].
- When planning the design of future studies based on the results of meta-analyses, *every effort should be made to explain between-study heterogeneity, using meta-regression or by obtaining and analyzing individual patient data* (whenever possible) [29]. In many cases, particularly in systematic reviews of diagnostic tests, unexplained between-study heterogeneity leads to very wide intervals around the predicted effects (e.g., test performance) in future studies. In the presence of substantial between-study heterogeneity, simply increasing the target sample size of planned studies cannot reduce the uncertainty of the predictive distribution because reducing sampling variability does not address heterogeneity. We did not conduct meta-regression or analyze individual patient data in our investigation; however, we do recommend such steps for planning the design of future studies based on meta-analytic results.
- We have assumed that studies provide unbiased estimates of test performance and probability of disease in their respective target populations. Although this is a common assumption in applied meta-analysis, we think that it is relatively implausible in most cases. *Our methods can be extended to account for (risk of) bias in the design and conduct of individual studies* [6,30,31]. In general, such adjustments result in even more plausible (greater) estimates of uncertainty around meta-analysis and decision analysis results.
- *Bayesian methods can be used for the analyses suggested previously because they provide a natural framework for evidence synthesis and prediction* and allow the incorporation of external information in a principled way. Nonetheless, classical (e.g., likelihood-based) methods for synthesis that do not require the specification of prior distributions, combined with forward Monte Carlo methods (sometimes referred to as probabilistic sensitivity analysis methods), can be used to perform analyses similar to those presented in this report [32].

Using an example from the diagnosis of breast cancer recurrence, we examined how alternative meta-analytic results can be used to populate decision models, with the goal of aiding the interpretation and contextualization of a meta-analysis of three diagnostic tests, specifically PET/CT, PET, and CW. In addition to offering the above recommendations as an extension of our analyses, we have demonstrated that using alternative summaries from an evidence synthesis, namely the random effects mean, the predictive estimate for a new study, and estimates of the true effect in specific studies can have an impact on clinical conclusions and their

precision. More concretely, our results show that, although PET/CT outperforms CW in a current setting (random effects mean) for the study populations included in this report, its effectiveness may not be greater than that of CW in a new setting (predictive estimate for a new study). In addition, in both current and future implementations, PET/CT may not outperform PET in diagnosing breast cancer recurrence in the study populations. This is especially evident in the future implementation scenario where the credible interval of the predictive estimate for the difference in QALYs between PET/CT and PET is considerably wider than that in the current setting scenario. Overall, we show that depending on the context in which the effectiveness of PET/CT, PET, and CW is evaluated, PET/CT may or may not outperform PET and CW. In this report, we considered effectiveness through diagnostic test performance and patient outcomes; however, decision analysts may wish to extend the notion of effectiveness to include cost and thus evaluate diagnostic tests through a cost-effectiveness perspective as well.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2019.01.002.

References

[1] Trikalinos T, Kulasingam S, Lawrence W. Deciding whether to complement a systematic review of medical tests with decision modeling. J Gen Intern Med 2012;27:76−82.

[2] Trikalinos T, Siebert U, Lau J. Decision-analytic modeling to evaluate benefits and harms of medical tests: uses and limitations. Med Decis Making 2009;29:E22−9.

[3] Pennant M, Takwoingi Y, Pennant L, Davenport C, Fry-Smith A, Eisinga A, et al. A systematic review of positron emission tomography (PET) and positron emission tomography/computed tomography (PET/CT) for the diagnosis of breast cancer recurrence. Health Technol Assess 2010;14:1−103.

[4] Auguste P, Barton P, Hyde C, Roberts T. An economic evaluation of positron emission tomography (PET) and positron emission tomography/computed tomography (PET/CT) for the diagnosis of breast cancer recurrence. Health Technol Assess 2011;15:iii−iv. 1−54.

[5] Higgins J, Thompson S, Spiegelhalter D. A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A Stat Soc 2009;172(1): 137−59.

[6] Welton N, Soares M, Palmer S, Ades A, Harrison D, Shankar-Hari M, et al. Accounting for heterogeneity in relative treatment effects for use in cost-effectiveness models and value-of-information analyses. Med Decis Making 2015;35:608−21.

[7] Dahabreh I, Chung M, Kitsios G, Terasawa T, Raman G, Tatsioni A, et al. Survey of the methods and reporting practices in published meta-analyses of test performance: 1987-2009. Res Synth Methods 2013;4(3):242−55.

[8] Trikalinos TA, Hoaglin DC, Small KM, Terrin N, Schmid CH. Methods for the joint meta-analysis of multiple tests. Res Synth Methods 2014;5(4):294−312.

[9] Chu H, Nie L, Cole S, Poole C. Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: alternative parameterizations and model selection. Stat Med 2009;28:2384−99.

[10] Ades A, Lu G, Higgins J. The interpretation of random-effects meta-analysis in decision models. Med Decis Making 2005;25:646−54.

[11] Spiegelhalter D, Abrams K, Myles J. Bayesian approaches to clinical trials and health-care evaluation. Chichester, West Sussex: Wiley; 2004.

[12] Raudenbush S, Bryk A. Hierarchical linear models: applications and data analysis methods. Thousand Oaks, California: Sage; 2002.

[13] Trikalinos T, Russell L, Sanders G. Evidence synthesis. In: Cost-effectiveness in health and medicine. New York, NY: Oxford University Press; 2016.

[14] Cooper N, Sutton A, Abrams K, Turner D, Wailoo A. Comprehensive decision analytical modelling in economic evaluation: a Bayesian approach. Health Econ 2004;13:203−26.

[15] Herbert H, Francis J, Rosenthal G. Cost effectiveness analysis applied to the treatment of chronic renal disease. Med Care 1968;6:48−54.

[16] Oppe M, Al M, Molken M-v. Comparing methods of data synthesis: re-estimating parameters of an existing probabilistic cost-effectiveness model. Pharmacoeconomics 2011;29:239−50.

[17] Vemer P, Al M, Oppe M, Molken MV. A choice that matters? Simulation study on the impact of direct meta-analysis methods on health economic outcomes. Pharmacoeconomics 2013;31:719−30.

[18] Teljeur C, O'Neill M, Moran P, Murphy L, Harrington P, Ryan M, et al. Using prediction intervals from random-effects meta-analyses in an economic model. Int J Technol Assess Health Care 2014; 30(1):44−9.

[19] Bossuyt P, McCaffery K. Additional patient outcomes and pathways in evaluations of testing. Med Decis Making 2009;29:E30−8.

[20] Jarvik J. Fundamentals of clinical research for radiologists: the research framework. Am J Roentgenol 2001;176(4):873−8.

[21] Thornbury J. Eugene W. Caldwell Lecture. Clinical efficacy of diagnostic imaging: love it or leave it. AJR Am J Roentgenol 1994;162(1):1−8.

[22] Mulherin S, Miller W. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med 2002;137:598−602.

[23] Begg C. Biases in the assessment of diagnostic tests. Stat Med 1987; 5:411−23.

[24] Walter S, Macaskill P, Lord S, Irwig L. Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. Stat Med 2012;31:1129−38.

[25] Whiting P, Rutjes A, Reitsma J, Glas A, Bossuyt P, Kleijnen J. Ources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140:189−202.

[26] Leeflang M, Bossuyt P, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol 2009;62:5−12.

[27] Graham P, Moran J. Robust meta-analytic conclusions mandate the provision of prediction intervals in meta-analysis summaries. J Clin Epidemiol 2012;65:503−10.

[28] Riley R, Higgins J, Deeks J. Interpretation of random effects meta-analyses. BMJ 2011;342:d549.

[29] Hinchliffe S, Crowther M, Phillips R, Sutton A. Using meta-analysis to inform the design of subsequent studies of diagnostic test accuracy. Res Synth Methods 2012;4(2):156−68.

[30] Greenland S. Multiple-bias modelling for analysis of observational data. J R Stat Soc Ser A Stat Soc 2005;168:267−306.

[31] Turner R, Spiegelhalter D, Smith G, Thompson S. Bias modelling in evidence synthesis. J R Stat Soc Ser A Stat Soc 2009;172(1): 21−47.

[32] Dias S, Sutton A, Welton N, Ades A. Evidence synthesis for decision making: embedding evidence synthesis in probabilistic cost-effectiveness analysis. Med Decis Making 2013;33: 671−8.