Original Article

Research Synthesis Methods

Received 19 September 2013, Revised 11 April 2014, Accepted 21 April 2014 Published online 3 June 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jrsm.1122

Combining randomized and non-randomized evidence in clinical research: a review of methods and applications

Pablo E. Verde* and Christian Ohmann

Researchers may have multiple motivations for combining disparate pieces of evidence in a meta-analysis, such as generalizing experimental results or increasing the power to detect an effect that a single study is not able to detect. However, while in meta-analysis, the main question may be simple, the structure of evidence available to answer it may be complex. As a consequence, combining disparate pieces of evidence becomes a challenge. In this review, we cover statistical methods that have been used for the evidence-synthesis of different study types with the same outcome and similar interventions. For the methodological review, a literature retrieval in the area of generalized evidence-synthesis was performed, and publications were identified, assessed, grouped and classified. Furthermore real applications of these methods in medicine were identified and described. For these approaches, 39 real clinical applications could be identified. A new classification of methods is provided, which takes into account: the inferential approach, the bias modeling, the hierarchical structure, and the use of graphical modeling. We conclude with a discussion of pros and cons of our approach and give some practical advice. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: observational studies; randomized control trials; bias modeling; network meta-analysis; cross-design synthesis; generalized evidence synthesis; hierarchical Bayesian models

1. Introduction

Statistical evidence synthesis is a branch of statistical methods that allows researchers to combine scientific results from multiple pieces of evidence into a single analysis. These techniques are used to extend the scope of a single experiment, by combining results from several experiments. A typical application is the meta-analysis of experiments addressing the same primary research question and using the same statistical design, where the application of simple statistical procedures (e.g., random-effects meta-analysis) is sufficient.

However, while in a meta-analysis, the clinical question may be simple, (e.g., what is the effect of an intervention in a population of interest?), the structure of evidence available to answer it may be complex (e.g., the published results may have different grades of quality), as a consequence, combining disparate pieces of evidence becomes a challenge.

In this review, we cover statistical methods that have been used for the evidence-synthesis of *different study types* with *the same outcome* and *similar interventions*. The study types considered are randomized controlled trials (RCTs) and non-randomized studies, covering studies with non-randomized control groups and studies without a control group (e.g., register and cohort study) (Deeks *et al*., 2003). The methods reviewed are methods used for combining aggregated data and used for combining aggregated with individual data as well.

The main reason for the aforementioned restrictions is the increasing complexity and quantity of methodological research in evidence-synthesis, which requires focusing on a methodological review. However, there is a strong need for such a specific review; for example, to answer the relevant question of how to generalize results from RCTs to clinical practice, that is, how much of the proved *efficacy* can be translated into *effectiveness*.

Coordination Center for Clinical Trials, University of Duesseldorf, Germany
**Correspondence to: Pablo E. Verde, Coordination Center for Clinical Trials University of Duesseldorf, Germany.*
E-mail: pabloemilio.verde@uni-duesseldorf.de

Moreover, researchers may have multiple motivations for combining different study types in a meta-analysis, for example:

- To increase the power to detect an effect that a single source of data is not able to detect.
- To reconstruct evidence that is not directly observable in a single study.
- To learn from the evidence how to improve the statistical design of future studies.
- To make decisions in situations where further experimentation may not be helpful, could not be ethical, or may not be feasible due to time or budget constraints.

However, no study type is free of bias, and the resulting analysis will be a trade-off between extending the inferential scope of a meta-analysis and adjusting the bias that is introduced by combining different study types.

The interest of including non-randomized studies in evidence synthesis has recently been highlighted in a special issue of this journal, where the authors presented the outcomes of a special workshop led by the Non-Randomized Studies Methods Group of the Cochrane Collaboration (Reeves *et al*., 2013). Four discussion papers covered the following: issues in study design and risk of bias by Higgins *et al*. (2013), issues relating to confounding factors when including non-randomized evidence by Valentine and Thompson (2013), issues in selective reporting by Norris *et al*. (2013), applicability of non-randomized evidence as complementary source of evidence by Schünemann *et al*. (2013), and a guideline of checklists for review authors by Wells *et al*. (2013).

Much has been written in evidence synthesis and meta-analysis from many perspectives. An early review of Bayesian meta-analysis methods in tutorial style is presented by Sutton and Abrams (2001). Probably, the most complete review in multi-parameter evidence synthesis is given by Ades and Sutton (2006). Sutton and Higgins (2008) presented an extensive review of methodological developments in meta-analysis. The paper of Higgins *et al*. (2009) concentrates on issues and applications of random-effects meta-analysis. Ioannidis (2010) reviews issues in meta-analysis from the practitioner's point of view.

This review updates previous methodological reviews (Sutton and Abrams, 2001; Ades and Sutton, 2006; Sutton and Higgins, 2008) in specific topics and includes new methods that were not developed at that time. Furthermore, a new classification of methods was developed and, for the first time, real medical applications of the methods assessed.

We omitted the highly important topic of publication bias, which addresses the problem that studies which claim statistically significant results are more likely to be published than studies with inconclusive results. Useful literature relating to this topic includes the following: Sutton *et al*. (2000), Rothstein *et al*. (2005) and the recent work of Copas (2013).

This paper is organized as follows: Section 2 describes the searching and classification techniques used to identify methodological work and their applications. Results of the methodological work are organized in chronological order and we provide an annotated description of the methods and their applications. Section 3 presents our results, and Section 4 provides a general discussion with some recommendations for practitioners.

2. Methods

2.1. Identification and classification of methodological work

Methodological papers have been previously identified in general reviews such as (Ades and Sutton, 2006) and (Sutton and Higgins, 2008). We use these reviews as a starting point to update the main methodological work and group them into methods, which investigated the combination of different study types in meta-analysis.

A manual search was performed by carefully looking at cross-references and in main applied statistical journals with a focus on applications in life sciences and medicine. Those included the following: Biometrics, Biometrical Journal, Biostatistics, Journal of the Royal Statistical Society series A and C, Research Synthesis Methods, Statistics in Medicine, and Statistical Methods in Medical Research. We also included main methodological journals, which publish applied work, those are the following: Annals of Applied Statistics, Journal of the American Statistical Association, Journal of the Royal Statistical Society series B, and Statistical Science.

For historical reasons, we start by presenting the Confidence Profile Method (CPM) in Section 3.1. Network meta-analysis is the topic of Section 3.2. The cross-design synthesis (CDS) and related approaches are covered in Section 3.3. Bias modeling of different study types is covered in Section 3.4, and the state of the art of Bayesian hierarchical models (BMHs) is presented in Section 3.5. In each section, we add a subsection with applications in clinical context. Section 3.6 summarizes the inferential approaches and operational characteristics of the statistical methods reviewed.

2.2. Identification of applications in clinical context

Clinical applications were identified from PubMed and within the Web of Science (Version 5.10), by using the following strategy:

- First, we select a key methodological paper in which the proposed method has been originally presented.
- Second, citations of the key methodological paper were identified and classified as follows:

- Methodological reference: In this case, the method is cited in a methodological or discussion context, where the method itself is not applied, but it is used as a reference for methodological extensions or discussion.
- Application in methodological context: This type of application is used for demonstration in a real data problem and used for methodological motivation or to highlight potential benefits in clinical use.
- Application in clinical context: Examples of clinical applications include the use of a method to provide scientific evidence for a clinical problem, the development of guidelines, or systematic reviews.

The citations databases used, with default starting date, were the following: Science Citation Index Expanded – 1945–present, Social Sciences Citation Index – 1956–present, Arts and Humanities Citation Index – 1975–present, Conference Proceedings Citation Index-Science – 1990–present, Conference Proceedings Citation Index-Social Sciences and Humanities – 1990–present, Book Citation Indexâ€" Science – 2005–present, Book Citation Indexâ€" Social Sciences & Humanities – 2005–present.

In addition, a PubMed search was performed with different search patterns: 'Confidence Profile Methods', 'Network meta-analysis', 'Cross-design Synthesis', 'Bayesian hierarchical model'; in combination with 'different study design' or 'meta-analysis'.

3. Results

3.1. The Confidence Profile Method

The CPM was introduced by Eddy (1989) as a general statistical framework to combine multiple sources of information in evidence synthesis and further described in a series of tutorial articles (Eddy, *et al*., 1992; Eddy, 1989; Eddy *et al*., 1990b; Eddy *et al*., 1990a; Shachter *et al*., 1990) and in a book with numerous examples (Eddy *et al*. (1992). The CPM was proposed under the realistic assumption that the empirical evidence used in meta-analysis could be incomplete, indirect, and biased.

The CPM was a vanguard approach, and it has influenced further developments over the last decades, including indirect treatment comparisons and network meta-analysis (NMA) (Section 3.2), direct bias modeling (Section 3.4) and the use of Bayesian graphical models in evidence syntheses (Ades, 2003; Spiegelhalter *et al*., 2004; Ades and Sutton, 2006) (Section 3.5).

Several important aspects have been introduced in the CPM framework, and we can highlight the following:

- First, the evidence to be analyzed in a systematic review is not considered as a realization of a random sample. As a consequence, statistical techniques with roots in the analysis of a single experimental data could lead to misleading results.
- Second, the problem of analysis of clinical evidence is embedded in a formal probability model with a Bayesian network representation. That allows a pictorial representation of the pieces of evidence, parameters of interest, functional parameters, and bias modeling.
- Third, the analysis of evidence is explicitly subjective. The analyst has to formalize his/her current state of knowledge of the problem at hand and include this aspect into the statistical model. Although statistical computations of the CPM can be done with direct use of the likelihood function or using Bayesian techniques, the interpretation is always subjective.
- The CPM emphasizes a case-specific modeling approach, where variability and bias of multiple sources of evidences have to be assembled in a single model. That contrasts with the statistical procedural approach, such as meta-analysis using fixed or random effects, where one approach applies to every situation.

3.1.1. Type of evidence, bias modeling, and inference. The CPM classified different *types of evidence*, where the *type of evidence* defines the likelihood function for interpretation of experimental results at *face value*. The main source of classification is the experimental design ((Eddy *et al*., 1992), Chapter 5).

Given that no experimental design is free of bias, Eddy *et al*. (1992) [p. 66–68] classified the propensity of bias of different experimental designs, with two main types of bias:

- *Bias to internal validity*, which is composed of factors that cause the observed results to not reflect the effects of the intervention in the circumstances of investigation. Typical examples of these factors are confounding variables, loss to follow-up, patient-selection bias, and dilution bias.
- *Bias to external validity and comparability*, which are composed of factors that make differences between the circumstances of investigation and the circumstances of interest. Examples of bias to external validity are population bias and intensity bias.

The CPM was not a BHM like those reviewed in Section 3.5. Statistical inference was carried out by direct application of Bayesian methods, that is, by multiplying the likelihood functions of the model parameters by their priors. Multiple parameters were assumed independent a priori and conjugate or Jeffrey's priors were used for

these model parameters. The method allowed to calculate posteriors of functional parameters, for example, parameters of interest after adjusting by bias modeling. Computations were based on Normal approximations of the posterior distribution and were implemented in the book's companion software (FAST*PRO) of Eddy *et al*. (1992).

As a general statistical framework, the CPM was perfectly suited to modern Bayesian computation techniques and software, but it was developed prior to the Markov Chain Monte Carlo (MCMC) revolution in statistics. Clearly, the Bayesian graphical approach was one of the more complex parts of the methodology. Although the diagrams were usually simple in their final form, they were not easy to develop unless the practitioner was skilled in structuring conditional independence statements between model quantities. Probably, these issues have restricted its application. Spiegelhalter *et al*. (2004) (Chapter 8) showed straightforward implementation of CPM's ideas with BUGS software (Lunn *et al*., 2009) including Bayesian graphical models and computations using MCMC.

3.1.2. Applications in clinical context. Eighty-five citations of two key methodological papers from Eddy and another 11 references were identified in PubMed. These papers were evaluated with respect to clinical applications:

1. Web of Science: Citations of Eddy (1989) ($n = 45$)
2. Web of Science: Citations of Eddy *et al*. (1990a) ($n = 40$)
3. PubMed: Search pattern 'Confidence Profile Method' ($n = 11$)

3.1.2.1. Guidelines. The CPM was systematically applied in a series of clinical guidelines developed by the American Urological Association and published between 1987 and 2007 in clinical urological journals. These guidelines cover the management of invasive bladder cancer (Eddy, 1989; Smith *et al*., 1999; Hall *et al*., 2007), ureteral calculi (Segura *et al*., 1997), female stress urinary incontinence (Leach *et al*., 1997), organic erectile dysfunction (Montague *et al*., 1996), prostate cancer (Austenfeld *et al*., 1994), and staghorn calculi (Segura *et al*., 1994).

In these guidelines, CPM was used for evidence combination, including meta-analysis of comparable RCTs, of individual arms of RCTs and of individual arms from all studies regardless of study design. The analyses were performed with the Fast*Pro software (Eddy *et al*., 1992).

Two publications describe a guideline for detecting development dysplasia of the hip in children, published in 2000 (Lehmann *et al*., 2000; Pediatrics, 2000). The method used a combination of expert panel, decision modeling, and evidence synthesis. Summarizing evidence was performed across probabilities by the CPM. The calculation was done with the BUGS software (Lunn *et al*., 2009).

3.1.2.2. Meta-analyses. In 2009, a meta-analysis on ovarian preservation during chemotherapy was published in the Journal of Women's Health (Clowse *et al*., 2009). Two systematic reviews using the CPM to combine evidence of RCTs and cohort studies were published in 2003 in the Journal of Hepatology, one dealing with acute hepatitis C (Licata *et al*., 2003) and the other with chronic hepatitis B (Craxi *et al*., 2003). The probability of sudden death from rupture of intracranial aneurysms was calculated in a meta-analysis and published in 2002 in the Journal Neurosurgery (Huang and van Gelder, 2002). In 1999, several meta-analyses were published, with meta-analytic techniques based on the CPM and using the FAST*PRO software. These studies covered hormone replacement therapy and the risk of colon cancer (Obstetrics and Gynecology, (Nanda *et al*., 1999), treatment of chronic hepatitis C (American Journal of Gastroenterology, (Leach *et al*., 1997) and Journal of Hepatology, (Craxi *et al*., 1999) and prophylactic auxiliary node dissection on breast cancer survival (Annals of Surgical Oncology, (Orr, 1999). Three applications of the CPM are related to meta-analysis in cardiology, one investigating predictors of adverse outcome after coronary interventions (Journal of American College of Cardiology, 1998, (Block *et al*., 1998) and two dedicated to risk stratification after myocardial infarction (Annals of Internal Medicine, 1997, (Peterson *et al*., 1997) American Journal of Cardiology, (Shaw *et al*., 1996).

In another application, CPM was used to derive a summary estimate of relative risk of future fractures from different study types, such as prospective cohort, case-control, and cross-sectional studies (Klotzbuecher *et al*., 2000). A systematic review of efficacy of ketogenic diet for the treatment of refractory epilepsy in children combining uncontrolled retrospective and prospective studies was performed with FAST*PRO (Lefevre and Aronson, 2000). The effect of spinal manipulation on patient's pain and functional outcomes in low back pain was assessed by combining data from 25 controlled trials (Shekelle *et al*., 1992). CPM was used to combine data from uncontrolled non-randomized trials into single best estimates of outcome of femoropopliteal percutaneous transluminal angioplasty in the treatment of lower extremity ischemia (Adar *et al*., 1989). Two meta-analyses, with a reference to CPM but combining only RCTs were performed, one dealing with antibiotics in tube thoracostoma (Evans *et al*., 1995) and the other with manipulation and mobilization of the cervical spine (Hurwitz *et al*., 1996).

3.2. Network meta-analysis

Network meta-analysis (NMA) is a new area in evidence synthesis, where the aim is to combine data from studies reporting randomized results of several treatments to not only make pairwise treatment comparisons, but to also reconstruct comparisons that have not been performed head-to-head in any study before.

Increasing interest of practitioners in this area has been recently surveyed by Abdelhamid *et al*. (2012). They reported that '…many reviewers (76%) accepted that indirect evidence is needed as it may be the only source

of information for relative effectiveness of competing interventions, provided that review authors and readers are conscious of its limitations'.

Network meta-analysis has its roots in the Eddy's CPM ((Eddy *et al*., 1992), p. 45), where disparate pieces of evidence are combined to reconstruct evidence, which is not directly observable. Early applications of indirect treatment comparison can be found in Higgins and Whitehead, 1996; Hasselblad, 1998; Dominici *et al*., 1999 and Ades, 2003 (see Section 4).

Meta-analysis, which combines results from a mixture of randomized treatment comparisons, has been called in different ways in the statistical literature: *mixed treatment comparisons* (Lu and Ades, 2004), *network meta-analysis* (Lumley, 2002), and *multiple-treatment meta-analysis* (Salanti *et al*., 2008). Statistical methods are based on the use of generalized linear modeling framework from the Bayesian (Dias *et al*., 2013) and classical perspective (Lu *et al*., 2012; Piepho *et al*., 2012). White *et al*. (2012) showed that NMA models can be estimated by expressing them as multivariate random-effect meta-regression models. Meta-analysis of aggregated and patient individual data has been investigated by exploring treatment by patient-level covariates interactions (Donegan *et al*., 2012; Donegan *et al*., 2013).

At first sight, statistical methods in NMA might be similar to the classical topic of incomplete block designs ((Hinkelmann and Kempthorne, 1994), sec. 9.8), where the number of experimental units in a block is smaller than the number of treatments. However, as pointed out by (Senn *et al*., 2011), while the randomization of treatments in incomplete block designs might be performed within and between blocks and the experimenter controls the distribution of treatments per block, in NMA randomization is only performed within the trial and experimenters do not have any control on the number of treatments per study. These issues make difficult to justify a valid measure of treatment effects at both the level of the study and across studies. Moreover, as usual in meta-analysis, trials are performed by different investigators on different patients and with different protocols. As a consequence, variability of treatment effects might be very different within and between trials. Therefore, modeling between-trial heterogeneity is not straightforward (Lu and Ades, 2009) and remains a modeling issue (Thorlund *et al*., 2013).

Another important issue that might arise in NMA is the lack of agreement between direct treatment comparison and evidence of indirect comparison. This type of conflict of evidence results when treatment differences vary between types of trials. Lumley (2002) called this issue *incoherence* while Lu and Ades (2006) called it *inconsistency*. Different statistical techniques have been proposed to detect and model inconsistency in NMA. Lu and Ades (2006) proposed a factor that measures inconsistency between treatment comparisons, while Dias *et al*. (2010) proposed a node-splitting algorithm to test inconsistency. For multi-arm NMA, Higgins *et al*. (2012) distinguish between two types of inconsistencies: *loop inconsistencies* and *design inconsistencies*. Loop inconsistencies are regarded as a special type of *between-studies heterogeneity* that might affect the magnitude of treatment effect. For example, studies of different comparisons were undertaken in different settings or contexts, and these differences are associated with the magnitude of treatment effect. Design inconsistencies are regarded as a study-level covariate that modifies the effect sizes *within the study*. They proposed an approach to identify inconsistencies by including a full set of design-by-treatment interaction terms in an NMA model. This model handles simultaneously design and loop inconsistencies.

Although, statistical methods of NMA of RCTs are an active area of research, the combination of studies with randomized and non-randomized evidence is a new area of research. We found two recent works on NMA and different study types: Schmitz *et al*., (2013) and Soares *et al*., (2014).

Schmitz *et al*. (2013) proposed three alternative approaches of combining data from different trial designs in NMA: a simple combination of study's data by ignoring the different design types; the usage of observational data as prior information to adjust for bias due to trial design; and a three-level hierarchical model to account for heterogeneity between-trial design. The first approach is used to analyze inconsistencies between direct and indirect treatment comparison. The second one is used to understand the bias that observational data may introduce into the analysis. This is performed with a prior to posterior sensitivity analysis. The third approach, the three-level hierarchical model, is used to combine different study types and to provide overall estimates after accounting for between-study type variability. This model is an application of the *grouped random-effects approach* that is reviewed in Section 3.5.1.

Soares *et al*. (2014) developed a hierarchical Bayesian model to include randomized and non-randomized studies in a NMA. Observational studies are used to explore modeling assumptions in evidence synthesis in the presence of sparse data.

3.2.1. Applications in clinical context.

1. Web of Science: Citation of Schmitz *et al*. (2013) ($n = 1$)
2. Web of Science: Citations of Soares *et al*. (2014) ($n = 0$)
3. PubMed: Search pattern 'network meta-analysis and different study types' ($n = 24$). Two clinical applications were identified.

The work of Schmitz *et al*. (2013) was cited by Mesgarpour *et al*. (2013), who combined 48 studies (34 RCTs and 14 observational) to compare safety of off-label erythropoiesis stimulating agents (ESAs) in critically ill patients. ESAs treatment is compared with other effective interventions, placebo or no treatment by using a three-level hierarchical Bayesian model. The model used by the authors accounted for between-studies variability and

between-design variability. In addition, a sensitivity analysis is performed by down-weighting the evidence of observational studies. They also analyzed the robustness of their results by comparing results from models that included all studies, with the results from models that excluded studies with high risk of bias or low quality. The authors concluded that there was no statistical evidence of increase of risk in ill patients treated with ESAs.

Bittl *et al*. (2013) performed a Bayesian cross-design and NMA of 12 studies (four randomized clinical trials and eight observational studies) comparing coronary artery bypass graft with percutaneous coronary intervention and seven studies (two randomized clinical trials and five observational studies) coronary artery bypass graft with medical therapy. Based on an NMA, they arrived to the conclusion that medical therapy is associated with higher 1-year mortality than with the use of percutaneous coronary intervention for patients with unprotected left main coronary artery disease (odds ratio, 3.22; 95% credibility interval, 1.96–5.30).

Jones *et al*. (2013) made a systematic review to compare effectiveness of antiplatelet therapy, medical therapy, exercise, and endovascular and surgical revascularization in patients with peripheral artery disease. A meta-analysis of direct comparison was supplemented with an NMA. Evidences were available from 83 RCTs and four observational studies.

3.3. *Cross-design synthesis*

The CDS was a method designed in 1992 by the US General Accounting Office to combine experimental and non-experimental data. The method is described in Droitcour *et al*. (1993) and in Chelimsky *et al*. (1993).

Cross-design synthesis was developed to adjust the typical patients selection bias of the RCTs and generalize their results to populations that have not been included in RCT experimentation. With this end in mind, historical information coming from registers should be combined with RCT's results. Under the CDS's paradigm, experimental and non-experimental data are viewed as complementary sources of information.

The typical application of the CDS is called the *empty cell problem*, where the results of the RCTs should be extrapolated to a subgroup population where the data are only available in the register. Basically, the logic behind the CDS is a step-wise strategy for evidence synthesis:

1. Assemble the literature on the effectiveness of an intervention and the individual patients data of subgroups of interest.
2. Determine whether bias is relevant in individual studies through expert review and then adjust for this bias (e.g., by using the CPM). Adjust the bias of individual data by covariates adjustment, standardization, propensity scores (Agostino, 1998), etc.
3. Combine the experimental and adjusted non-experimental evidence by assuming that clinical effects are proportional between subgroups.

The reliability of the CDS was criticized in the Lancet by Anonymous (1992) and by Begg (1992) who pointed out that the authors have underestimated the problem of harmonizing results from RCTs and medical databases.

A modern view of the CDS was recently given by Kaizar (2011), where the statistical framework proposed by Imai *et al*. (2008) is used to evaluate the statistical properties of a CDS estimator. Kaizar (2011) evaluated the CDS with an extensive computer simulation experiment and used a real case example regarding the effectiveness of insulin pumps versus glargine insulin injections in the regulation of blood glucose in adolescents with type 1 diabetes. (Kaizar, 2011) concludes that under reasonable data assumptions, the simple CDS estimator has smaller bias and better coverage than commonly used estimates based on randomized or observational studies alone.

Another topic directly related to the aims of the CDS is the assessment of effectiveness, that is, the generalization of RCTs results to clinical practice. RCTs provide the gold standard for proving efficacy of interventions. The reason is high internal validity, allowing causal reasoning. Often, RCTs are performed with highly selected patient populations, excluding women, children, elderly, and patients with comorbidity. As a consequence, generalizability of results from RCTs to these patients is severely limited, and different types of bias, such as selection bias, may occur. In addition, adequate information about the recruitment process is often not provided, making an assessment of generalizability to clinical practice difficult. Recent work in this area is presented by Benson and Hartz (2000), Zimmerman *et al*. (2004), Fortin *et al*. (2006), Prentice *et al*. (2006), Greenhouse *et al*. (2008), Ahern *et al*. (2009), Frangakis (2009), and Cole and Stuart (2010).

Some developments in BHMs have been motivated by the CDS method, for example, those by Nixon and Duffy (2002), Prevost *et al*. (2000), and Peters *et al*. (2005). We review these methods in Section 3.5.

3.3.1. *Applications in clinical context.*

1. Web of Science: Citation of (Droitcour *et al*., 1993) ($n = 9$)
2. Web of Science: Citations of (Chelimsky *et al*., 1993) ($n = 3$)
3. PubMed: Search pattern 'cross-design synthesis' ($n = 9$)

The CDS method was applied by the US General Accounting Office in 1994 to study the effect of 'Breast conservation versus mastectomy: patient survival in day-to-day medical practice in randomized studies'. No other clinical application was found.

3.4. Direct modeling of bias

3.4.1. Classical meta-analysis techniques. Meta-analysis with historical controls is analyzed by Begg and Pilote (1991). A random-effects meta-analysis model is presented in which the baseline effect in each study is random, but the treatment effect is constant. With this model, the appropriate contribution of historical studies can be determined. The method is Bayesian in nature, but estimation of hyper-parameters is performed by Empirical Bayesian techniques. The authors illustrated this method by combining four RCTs with 12 uncontrolled studies to analyze the efficacy of bone-marrow transplantation versus conventional chemotherapy in the treatment of acute non-lymphocytic leukemia. Li and Begg (1994) presented a non-iterative estimator of treatment effects based on this method. They studied theoretical properties and presented results from a simulation experiment which contemplate different random-effects distributions (normal, log-normal, exponential, and uniform). They concluded that both the pooled effects and the between-studies estimators are strongly consistent with desirable heuristic properties.

An early work in combining disparate study designs is presented by Brumback *et al*. (1999). They present a meta-analysis where three case-control, and 28 cohort studies are combined to study the association of prenatal testing via chorionic villus sampling with the occurrence of terminal limb defects. The authors combine two types of sub-models in a single meta-analysis: a fixed effect sub-model with a logistic-regression is used to model the evidence of case-control studies and a random-effects with a Poisson regression with a conjugate Gamma distribution is used to model the evidence of the cohort studies. Inference on the pooled effect parameter is estimated by combining the likelihood of the fixed effect, and the marginal likelihood of the random-effects model. The resulting likelihood depends on the parameters of the Gamma distribution, the authors presented a sensitivity analysis by estimating the pooled effect for different values of these parameters.

3.4.1.1. Applications in clinical context. 1. Web of Science: citations of Begg and Pilote (1991) ($n = 25$). In Web of Science, we found 25 methodological citations of Begg and Pilote (1991). No clinical applications were found using the random-effects model of Begg and Pilote (1991).

3.4.2. Adjustment of likelihoods for study design and quality. Wolpert and Mengersen (2004) presented reductionist and alternative method to the CPM. While CPM constructs a global probabilistic model by using conditional independence between model parameters and pieces of evidence, Wolpert and Mengersen (2004) proposed to directly adjust the likelihood of each study's parameter for its potential bias. The adjustment is done by defining a bias function similarly to the CPM. In a second step, the adjusted likelihoods are combined by a meta-analysis model (e.g., a random-effect model). Computations were based on MCMC, and the authors highlight potential advantages of the method, such as inference of functional parameters and ranking parameters.

They apply this technique to combine case-control studies with cohort studies in order to assess the relationship between environmental exposure to tobacco smoke and lung cancer. The likelihood of each study is adjusted by the propensity that each design has with respect to different types of misclassifications, which includes the following: the bias introduced from the misclassification of people who always smoked to people who have never smoked, bias of misclassification of disease and non-disease, and misclassification of exposure status.

Multiple bias modeling of meta-analysis of retrospective case-control studies is analyzed by Greenland (2005). He presented a Bayesian modeling approach where priors are used to encapsulate external information of different types of bias. He called this sort of sensitivity analysis a meta-sensitivity modeling. He applied this technique to adjust a meta-analysis of 14 case-control studies (12 published and two unpublished) of residential magnetic fields and childhood leukemia. The sources of bias considered in the meta-sensitivity analysis were the following: confounding factors of field exposure and leukemia, sampling and response bias, and measurement errors in magnetic fields.

Another way to adjust likelihoods in Bayesian modeling is by explicitly discounting for study's quality bias using a 'power prior' (Ibrahim and Chen, 2000). The likelihood of low quality studies is raised to a power factor between 0 and 1, where values close to 0 indicate low quality and values close to 1 no bias. Recently, Neuenschwander *et al*. (2009) proposed to scale the power priors to a 'proper power prior' to estimate the discounting factor.

Turner *et al*. (2009) recognize the practical limitations and difficulties of elicitation of bias, and they introduced a comprehensive approach to adjust a classical meta-analysis for multiple sources of bias. The idea is that the pooled treatment effect of the bias-adjusted meta-analysis will reflect a more realistic estimate than the naive meta-analysis.

In Turner's approach, multiple sources of bias are divided into two main types of bias: internal validity bias and external validity bias. Each study included in the meta-analysis is evaluated by a group of assessors, who estimate different types of biases by a score system. External empirical evidence of bias can be included in the analysis (Welton *et al*., 2009), but the method assumes that, in general, it is unrealistic that such evidence exists.

3.4.2.1. Applications in clinical context. 1. Web of Science: citations of Wolpert and Mengersen (2004) ($n = 14$). We found 14 citations of Wolpert and Mengersen (2004) with three clear related clinical applications:

- Bayesian modeling for direct adjustment of likelihoods was applied to the outcome of beta-interferon treatment in relapsing-remitting multiple sclerosis (O'Rourke *et al*., 2007). In this analysis, the likelihood of the log odds ratio is approximated by a normal distribution, the results from observational case-control studies are adjusted to account for an exaggerated precision of treatment effect and for a systematic bias toward overestimation of treatment effects. The adjustment is based on a fixed value, which reflects that observational studies overestimate treatment effect by a median of 30% (Egger *et al*., 2002).
- A similar approach was presented by O'Rourke *et al*. (2009), where safety and efficacy of IV-TPA for ischaemic stroke were analyzed by a cumulative Bayesian meta-analysis based on a Beta-Binomial model. In this case, results from observational studies are adjusted by overestimation of treatment effect by using a fixed approach.
- Another cumulative meta-analysis with bias adjustment for observational studies is presented by O'Rourke and Walsh (2010). A prior distribution for the OR of dead within 1 year after acute stroke was built using a meta-analysis of 26 RCTs comparing stroke unit care versus alternative models of stroke care. The analysis was performed sequentially by starting with the RCTs prior, data from individual observational studies were used to sequentially update outcome knowledge. Again, the likelihood of each observational study was adjusted for overestimation by using a fix value, which reflects that observational studies overestimate treatment effect by a median of 30% (Egger *et al*., 2002).

The assertion in the examples seems to be that the non-randomized studies overestimate the effect size, whereas Deeks *et al*. (2003) clearly demonstrated that non-random allocation can lead to overestimation or underestimation of treatment effects.

Recently, Turner *et al*. (2012) applied their method to adjust a meta-analysis, which included 10 studies comparing routine antenatal anti-D prophylaxis to control. After adjustment for differences in study design and quality, the authors concluded that there is strong evidence in the benefit of routine antenatal anti-D prophylaxis.

3.5. Bayesian hierarchical methods

Bayesian hierarchical modeling techniques have been used to combine studies with different designs during the last two decades. In this section, we consider full BHMs where uncertainty of the hyper parameters are included into the model and where computations based on MCMC are used to estimate posteriors of all parameters in the model in a single modeling step.

3.5.1. The grouped random-effects approach. Combining dissimilar studies in a common meta-analysis was criticized by Larose and Dey (1997). They proposed to group studies with different designs in a common BHM, where each group has its own treatment effect and dispersion parameter. They called this approach 'the grouped random-effects' model. They illustrated their technique with a meta-analysis which combined six single-blind RCTs with nine double-blind RCTs in the study of efficacy of an anti-epileptic drug, progabide. The model is a binomial-normal BHM where a careful sensitivity analysis of the hyper-priors is analyzed. The authors presented four non-informative models for the hyper-priors. Computations were implemented by using Gibbs sampling and the Metropolis method. They concluded that results were insensitive to the hyper-priors specification. Interesting results of this analysis were that open studies were systematically more dispersed than closed studies, and open studies supported the efficacy of progabide, closed studies supported the reverse hypothesis, while the union of the groups supported neither hypothesis. That was a clear warning for meta-analyses that indiscriminately combined studies with different designs.

Prevost *et al*. (2000) presented the first formal Bayesian approach to the cross-design synthesis problem. They propose a three-level hierarchical model, where the first and the second level are used to model the observed evidence and the variability between studies, respectively. A third level is used to model the variability between-study types. This model allows the exchange of information across the study types, with the additional advantages that neither assumes independence between effects in different study types nor equivalence of such effects. In addition, the authors describe a posterior predictive analysis to the 'empty cell' problem, where results of a new RCT or a non-randomized study are predicted from the model. The model is illustrated by combining evidence of RCTs and non-randomized studies, which describe the benefit, in terms of mortality reduction, of using mammography screening in breast cancer for different age groups of women.

Prevost *et al*. (2000) described carefully how priors for hyper-parameters were chosen, and they presented a sensitivity analysis for the priors specification. They concluded that the variability between-study types has the greatest effect on both, the estimate of the overall pooled effect, and the pooled effects within each type of studies.

Another Bayesian development of the cross-design synthesis is presented by Peters *et al*. (2005). The model is motivated by a toxicological application, which investigated the association between exposure to trihalomethanes in drinking water and low birth weight. The available evidence included the study-specific dose-response slope from studies across two disciplines: epidemiological studies with evidence of humans and toxicological studies with evidence of animals. A three-level BHM is developed to account for study type effects, which is similar to the model of Prevost *et al*. (2000). The authors presented a detailed sensitivity analysis by using

different sets of prior distributions. They arrived at similar conclusions as Prevost *et al.* (2000), where priors on the between-study types variance component had a main influence in the analysis.

3.5.2. The hierarchical regression modeling approach. Another area is the combination of aggregated and patient individual data, where both types of evidence correspond to different study designs. Methods for combining aggregated and individual patient data have been developed recently under the name of *Hierarchical Related Regression* (HRR) modeling (Jackson *et al.*, 2006; Jackson *et al.*, 2008). The main idea of HRR is the existence of shared parameters between different data sources that justify merging information in a common model. In HRR, there is an explicit use of graphical models to describe: the probabilistic relationship of multiple sources of information, which bias sub-models are introduced and how share parameters are linked to different data types. Computations are usually implemented in WinBUGS or other MCMC software (e.g., OpenBUGS, JAGS). Recent applications and further development of HRR are presented by Molitor *et al.* (2009) and Jackson *et al.* (2009). Riley *et al.* (2008) and Sutton *et al.* (2008) described similar approaches of combining aggregated and individual data in meta-analysis of randomized trials.

McCarron *et al.* (2010) combined RCTs and non-randomized studies to syntheses evidence of studies comparing treatment for abdominal aortic aneurysms. They developed a BHM, where each arm's outcomes are modeled with binomial distributions, and study effects are modeled with a normal distribution in the logistic scale (i.e., $log(p/(1-p))$). Systematic variability between different study types are modeled by adjusting the study effects with a meta-regression model. The authors proposed to adjust differences in patients' characteristics between study arms. For example, if age is used for adjustment at the study level, the difference of age between study arms is used as covariate. The idea behind this type of adjustment comes from the empirical finding of Deeks *et al.* (2003), which describe that non-randomized trials tend to present unbalance in patients' characteristics between studies arms. The authors argued that covariate adjustment using aggregate study values does not account for covariate imbalances between treatment arms. In a complementary work, McCarron *et al.* (2011) presented an exhaustive simulation experiment to validate the idea of adjustment by differences between arms in patient characteristics.

3.5.3. The hierarchical weighting approach for study design and quality. Complex cost-effectiveness modeling is an area where evidence is usually collected from different study types. Spiegelhalter and Best (2003) embed a generalized evidence synthesis model into a cost-effectiveness model to predict costs and benefits of hip prostheses in different age-sex subgroups. They introduced a BHM for generalized evidence synthesis where multiple sources of evidence could be weighted according to their assumed quality. In this model, the study effect is the sum of two random effects: one describing the study's external bias and the other describing the study's internal bias. The marginal variance of study's effect is expressed as the product of study's quality weight and the variance between studies due to external bias. The quality weights are interpreted as the proportion of between-study variability unrelated to internal bias. This strategy avoids the estimation of the second variance component related to internal bias. For the quality weights, the authors proposed to give fixed values. These values can be obtained from external empirical information or by elicitation from expert opinion. In either case, a sensitivity analysis to a range of assumptions about the quality weights can be carried out. An example of combining one RCT, one register and one case series is used to illustrate this technique. A sensitivity analysis for different quality weight values is presented where the evidence of non-randomized evidence is down-weighted in different ways.

Welton *et al.* (2009) presented a BHM to model meta-analysis or RCTs that may present a high risk of bias. In particular, the authors consider RCTs that may be biased by failure to conceal randomized allocation at the time of patient recruitment. The authors developed a mixed effects model where treatment effects are considered as fixed and bias effect as random. One novelty of this work was to inject empirical bias information into the model by using prior distributions that are estimated from a collection of previously published meta-analysis of RCTs. Although this model is developed only for RCTs, it can be directly applied to combine experimental and non-experimental studies, where the last ones are at high risk of bias.

Meta-analysis of diagnostic tests is an area where RCTs are usually combined with observational studies. The main motivation is to assess diagnostic accuracy in populations that are not contemplated in RCTs. This type of meta-analysis required special techniques to model the correlation between test operating characteristics (e.g., sensitivity and specificity). Verde, (2010) developed a BHM, where random effects follow a bi-variate scale mixture distribution. He gave direct interpretation of the scale weights as measures of model's deviations. A systematic increase of dispersion of retrospective studies was modeled by allowing a meta-regression equation to the scale weights. This technique is illustrated in a meta-analysis of 51 studies, which investigate the accuracy of computer tomography in the diagnoses of appendicitis. The model is implemented in the R package *bamdit* (Bayesian meta-analysis of diagnostic test data) (Verde, 2013), which combines R and JAGS (Just another Gibbs sampling) (Plummer, 2003). The use of scale mixture distributions is a potential modeling tool to handle different study types in meta-analysis of efficacy outcomes as well.

3.5.4. Further hierarchical modeling techniques. Dominici *et al.* (1999) combined results of RCTs with heterogeneous designs to analyze the effectiveness of commonly recommended prophylactic treatments for migraine headaches. They developed a complex BHM to handle a diversity of reporting results (some studies reported results in continuous scores, others reported differences between treatments, others dichotomous outcomes) by using a latent variable approach. Studies presented different type of treatments and indirect comparison was also used to assess treatments that were not compared in the same trial. This work is one of the earliest full implementation of ideas coming from Eddy's CPM by using modern Bayesian modeling and computational techniques (e.g., MCMC). A related work on mixed treatment comparisons was presented by Ades (2003), who extended the ideas of Eddy's CPM on 'chain of evidence' to reconstruct treatment comparisons where no direct comparison evidence was available. The work of Dominici *et al.* (1999) and Ades (2003) represent an early development of mixed treatment comparisons in meta-analysis.

While cross-design synthesis refers to the inclusion in a meta-analysis of studies addressing the same question under different designs, Nixon and Duffy (2002) proposed to combine studies addressing different but clinically related questions. They called this procedure 'the cross-issue synthesis', which was another name for the 'chain of evidence' problem.

The authors build a BHM to estimate the effectiveness of tamoxifen in the treatment of breast cancer for women with mutations in the BRCA1 or BRCA2 gene. One factor affecting the effectiveness of tamoxifen is the estrogen-receptor (ER) concentration of the primary tumor. Women with this gene mutation are typically ER negative, so the effectiveness of tamoxifen is affected by this mutation. They estimate the effectiveness of tamoxifen in BRCA by combining three different study types: preventive trials of tamoxifen, studies of adjuvant tamoxifen and studies reporting relationship between ER gene mutations. The authors used the *grouped random-effect* approach to allow different variability parameters for each study type and functional parameters to reconstruct the conditional probabilities needed in the analysis. This analysis was an example of what we can call today "research synthesis for personalized medicine".

Meta-analysis is usually a two-step analysis: In the first step, individual studies are selected and summarized and in the second step a meta-analysis model is applied (e.g., a random-effects model). This contrasts to BHM where a single step is used to estimate all parameters simultaneously. The BHM approach has the advantage of contemplating all parameters' variability in a single model and it offers great technological flexibility by using MCMC methods. However, there are situations where the two-step approach is useful: when study-specific analyses are too complex, when there are several models or parameters of interest to consider or when the parameters of interest are complex functions of other study parameters. Recently, Lunn *et al.* (2013) presented a new strategy of meta-analysis, where a Bayesian two-step approach is proposed. The idea is to give a full Bayesian analysis at the level of each study and summarize study results by the posteriors resulting from MCMC. In the second step, parameters' posteriors for each study are combined in a global Bayesian meta-analysis model. The authors illustrate this new technique with two examples: one meta-analysis that studies the effect of taking diuretics on the risk of pre-eclampsia during pregnancy and another complex meta-analysis where studies provide longitudinal measures of abdominal aortic aneurysms data together with the occurrence of clinical events. Clearly, this new meta-analysis approach can be directly used for combining studies of different design, for example individual bias modeling can be applied to each study in the first stage and combination of study results in the second one.

3.5.4.1. Applications in clinical context.

1. Web of Science: Citations of Larose and Dey (1997) ($n = 20$).
2. Web of Science: Citations of Dominici *et al.* (1999) ($n = 20$).
3. Web of Science: Citations of Prevost *et al.* (2000) ($n = 44$).
4. PubMed: Search pattern 'Bayesian hierarchical model' in combination with 'different study design' ($n = 8$) or 'meta-analysis' ($n = 21$).

For Larose and Dey (1997) and Dominici *et al.* (1999) no clinical applications were found. From 44 citations of Prevost *et al.* (2000), two clinical applications were identified: one was a Bayesian meta-analysis from Grines *et al.* (2008), which compared short-term mortality estimates from RCTs and non-RCTs in the intervention of acute myocardial infarction using AngioJet thrombectomy to percutaneous coronary intervention alone. The other was a BHM from Sampath *et al.* (2007), which assessed the efficacy of loop diuretics in acute renal failure in a meta-analysis by combining RCTs and non-RCTs.

3.6. General characteristics of evidence synthesis methods

The previous sections were divided by the proportion of influence that classical and Bayesian methods have on the development of methods for combining different study types. However, no matter which statistical school, these methods have a particular characteristic: the necessity of bias modeling between pieces of evidence, which clearly introduce an overlapping area between techniques. The aim of this section is to provide a more general understanding of how those methods overlap in terms of the statistical philosophy and the bias modeling technique.

Table 1 represents a classification of statistical methods used in research synthesis. Methods are characterized according to the following features:

- Statistical inference: A method is classified as *Classical* or *Bayesian*, where *Bayesian* means that prior distributions for all parameters are given. For example, the commonly used random-effects model, where all parameters are estimated from the data (i.e., Empirical-Bayes estimation), is considered as a *Classical* inferential approach.
- Bias modeling: We classified the bias modeling as *Yes*, if *explicit modeling of bias* is used (e.g., quality weighting and likelihood adjustment).
- Hierarchical modeling: This feature is classified as *Yes*, if the method involves hierarchical parameter structures to model multiple sources of evidence.
- DAG: *Yes* means that the method is based on a Directed Acyclic Graph representation. DAGs representations were promoted in the early days by (Eddy, 1989) and it is interesting to assess if this feature has been used.

Starting at the top of Table 1, we have as a reference the most popular meta-analysis methods: the fixed-effects and the random-effects models. These methods are blind to potential bias, if they are used for combining different study types, their results are prone to a multiplicity of bias.

The Eddy's CPM is represented as a hierarchical Bayesian meta-analysis with the possibility of extensive bias modeling. Eddy's method was an attempt to improve the bias issues of fixed and random-effects models. However, it is interesting to note that the clinical applications we found used the CPM as a Bayesian random-effects meta-analysis without bias modeling. In some cases, the authors mentioned that the CPM could adjust the meta-analysis when different study types are combined, but they did not make bias adjustment themselves. Eddy himself laments that the complexity of his method has limited its use among practitioners (Eddy, 2013).

Following our historical approach, the cross-design synthesis is classified as a sort of classical meta-analysis with explicit modeling of bias. The work of Begg and Pilote (1991) as well as Brumback *et al*. (1999) uses classical statistical methods, with explicit bias modeling in the case of Brumback *et al*. (1999).

The rest of the papers in Table 1 clearly show that during the last 15 years, the Bayesian approach has dominated this area of meta-analysis. The *grouped random effects* approaches did not focus on bias modeling but on variability between study types, while the *direct adjustment of likelihoods* (Wolpert and Mengersen, 2004; Greenland, 2005; Turner *et al*., 2009), *the hierarchical regression* (Jackson *et al*., 2006; McCarron *et al*., 2010), and *weighting approaches* (Spiegelhalter and Best, 2003; Welton *et al*., 2009; Verde, 2010) have enforced bias modeling.

Table 1. General characteristics of evidence synthesis methods used to combine different study types.

Main reference/method	Statistical Inference	Bias modeling	Hierarchical	DAG
Fixed effects meta-analysis	Classical	No	No	No
Random-effects meta-analysis	Classical	No	Yes	No
Confidence profile method	Bayesian	Yes	Yes	Yes
Cross-design synthesis	Classical	Yes	No	No
(Begg and Pilote, 1991)	Classical	No	Yes	No
(Brumback *et al*., 1999)	Classical	Yes	Yes	No
(Wolpert and Mengersen, 2004)	Bayesian	Yes	Yes	No
(Greenland, 2005)	Bayesian	Yes	Yes	No
(Turner *et al*., 2009)	Bayesian	Yes	Yes	Yes
(Welton *et al*., 2009)	Bayesian	Yes	Yes	No
(Larose and Dey, 1997)	Bayesian	No	Yes	No
(Prevost *et al*., 2000)	Bayesian	No	Yes	No
(Peters *et al*., 2005)	Bayesian	No	Yes	No
(Jackson *et al*., 2006)	Bayesian	Yes	Yes	Yes
(Riley *et al*., 2008)	Classical	Yes	Yes	No
(Sutton and Higgins, 2008)	Bayesian	Yes	Yes	No
(McCarron *et al*., 2010)	Bayesian	Yes	Yes	No
(Spiegelhalter and Best, 2003)	Bayesian	Yes	Yes	Yes
(Verde, 2010)	Bayesian	Yes	Yes	No
(Dominici *et al*., 1999)	Bayesian	No	Yes	No
(Schmitz *et al*., 2013)	Bayesian	Yes	Yes	No
(Soares *et al*., 2014)	Bayesian	Yes	Yes	No

The aim of this summary is to show the relative influence of Bayesian, frequentist, and bias modeling upon different methods developed in the last two decades.

The recent work in combining randomized and observational studies in NMA can be clearly classified. Schmitz *et al*. (2013) is a three-level BHM based on the *grouped random effects* approach to model between-study type heterogeneity. The work of Soares *et al*. (2014) is a BHM with extensive bias modeling.

Finally, the use of DAGs in evidence synthesis has been sporadic and more related to the use of the statistical software (e.g., WinBUGS).

4. Discussion

4.1. Classification of statistical approaches

This paper aims to give an overview of different modeling techniques that have been developed to combine different study types in meta-analysis. For historical reasons, we started with the Eddy's confidence profile method and continued with the cross-design synthesis, but the classification between bias modeling and Bayesian hierarchical modeling was less clear.

The classification in Section 3.6 was an attempt to clarify the overlapping areas between those sections. Although, this is only a rough classification some clear patterns emerge: Independently from the inferential approach, bias modeling is promoted almost for every model, the increasing applications of hierarchical Bayesian modeling, with the classical techniques that have been wiped off the play field. The trend of case-specific modeling approach as was originally promoted by the CPM.

4.2. Critique of our review and the methodological impact on clinical applications

In this review, clinical applications mean applications of the methodology to improve diagnosis, prognosis, or treatment for a clinical problem. It can be assumed that the majority of serious clinical applications have been published in at least one of the databases covered in our review. Nevertheless, it was difficult to identify and link statistical methods to clinical applications. Our searching strategy for identification has been weak in many aspects including the following points: First, there is not always one methodological paper, which can be clearly defined as the origin source of a method/technology. Second, even if such a paper exists, it may not be cited in a clinical application. Therefore, our approach to look at citations may not find clinical applications. Third, a broad search in PubMed without specification gives too many publications (e.g., BHM: 1149). A more restricted search, for example, 'Bayesian hierarchical model' + 'meta analysis', may again miss clinical applications.

As a consequence, our strategy (citation of key papers and restricted search) may identify only a subset of the clinical applications. Nevertheless, it is a systematic and reproducible strategy and for the review more than 250 publications, which have been identified according to this strategy, have been evaluated, and only 39 clinical applications were found. However, from a pragmatic point of view, we can at least have a rough estimate of the amount of clinical applications.

Taking a historical perspective, the impact of methodological work in clinical applications can be summarized as follows: early ideas of the Eddy's CPM were adopted by research groups and guidelines were developed, but the method did not spread out in practice. Classical approaches like the one of Begg and Pilote (1991) and the cross-design synthesis were not applied in real clinical context. Adjusted likelihood techniques were applied by a research group, but they did not reach general practice. The potential that BHM has in complex meta-analysis modeling has been established with a large amount of examples, and methodological work but expertise required for their applications remains an issue.

4.3. Relationships with network-meta-analysis

Specific to NMA are inconsistencies resulting from differences in treatment effects across direct and indirect comparisons, which may result in bias. However, if there are sources of bias that effect direct comparisons of studies then the pooled results of NMA incorporating different study types (e.g., RCTs and studies with non-randomized control groups) are affected as well. In addition, if estimation of between-study heterogeneity is of major issue in NMA, the inclusion of different study types can challenge practitioners in this problem. Therefore, biases generated by combining different study types are also relevant for NMA. The recently work of Schmitz *et al*. (2013) and Soares *et al*. (2014) are two examples of the recent trend in this new area of research.

One potential advantage of combining observational and RCTs in NMA is that we might have direct treatment comparisons in observational studies that are not represented in the RCTs. Making available direct comparisons form observational studies might reduce the risk of having inconsistencies in the NMA, at the expenses of introducing bias due to non-randomized treatment assignment.

4.4. Influence of statistical software

The methodological development in this area has been strongly influenced by the statistical software BUGS (Lunn *et al*., 2009) and Bayesian methodological papers published after 2000 used BUGS. Moreover, the published BUGS

scripts allow practitioners to use this software in their own applications. This trend contrasts with papers published during the nineties, where the main focus was on methodological research with little chance of using these methods in clinical contexts. We can consider this development as a success of the early ideas of Eddy's CPM.

Compared with other statistical areas, the development of R packages for meta-analysis has been slow and simplistic where most of the R packages for meta-analysis are focused on single study type meta-analysis. There is a lot of work that remains to be done in software development in this area.

4.5. Some practical advice

Combining different study types in a single meta-analysis is motivated by the principle of using all of the available evidence in a meta-analysis. However, we have seen in this review that there are many alternative methods to perform this task. Some of these methods require substantive input from outside the statistical analysis (e.g., the Turner bias model).

Clearly, transparency in the data collection and detailed information on each study included in the review is one of the basic premises in meta-analysis, but combining different study types demand an extra modeling effort. We add the following advice to practitioners in this area:

- Regardless of which meta-analysis approach is used, we should investigate external sources of information that may help in the bias modeling process. We could use this information for prior elicitation of bias (Turner *et al*., 2009), for direct likelihood adjustments (Wolpert and Mengersen, 2004), for meta-regression approaches (McCarron *et al*., 2010), for empirical bias modeling (Welton *et al*., 2009), or for quality weighting (Spiegelhalter and Best, 2003).
- Before combining different study types in a single meta-analysis, we should first make a separate meta-analysis for each study type. Exploring the differences and contradictions between results may help the modeling process. For example, increases of variability between-study types may be resolved by using grouped random-effects techniques (Larose and Dey, 1997; Prevost *et al*., 2000).
- We may ask to which extent does the model fitted predict future results? Model validation in meta-analysis is not very popular, but it should be like any other statistical modeling problem. Bayesian predictive data are conditionally independent from the data used to build the model and can be used for model checking in meta-analysis (Higgins *et al*., 2009; Verde, 2010).
- Can we detect conflict between pieces of evidence? The *conflict assessment* is the *deconstructionist side of evidence synthesis*, where each piece of evidence is put aside from the full model and compared to the rest of the evidence. Conflict assessment of pieces of evidence in meta-analysis is a new area of methodological research. One possibility is to embed a meta-analysis model in a more general model where the non-conflict situation is a particular case. For example, Verde 2010a applied a scale mixture of multivariate normal and he made conflict diagnostics by direct interpretation of the scale weights. Another alternative is presented by Presanis *et al*. (2013), where the authors described how to generalize the conflict *p*-value proposed by Marshall and Spiegelhalter (2007) to complex evidence modeling.
- Unfortunately, bias modeling cannot be validated, but a sensitivity analysis based on predictive data can be used to understand how conclusions from a meta-analysis are affected by the inclusion of different study types. We should have in mind that usually there is not 'a best model'. Examples of applications such as those described by Spiegelhalter and Best 2003) show that combining disparate evidence ends in a stochastic sensitivity analysis and not to a single best analysis.
- Bayesian hierarchical models have been the most popular approach for combining disparate sources of evidence, but there are a number of issues from the practical perspective, such as when to judge studies or study types 'exchangeable', how to put suitable priors on variance components, which type of sensitivity analysis is particularly relevant, and so on.

5. Acknowledgements

The authors are very grateful to David Spiegelhalter for his comments on a draft version of this paper. We also would like to thank the associate editor for his comments and two reviewers (supplied anonymously), which helped us to improve the quality of this paper. This work was supported by the German Research Foundation project DFG Oh 39/11-1.

References

Abdelhamid A, Loke Y, Parekh-Bhurke S, Chen Y, Sutton A, Eastwood A, Holland R, Song F. 2012. Use of indirect comparison methods in systematic reviews: a survey of Cochrane review authors. *Research Synthesis Methods* **3**: 71–79.

Adar R, Critchfield G, Eddy D. 1989. A confidence profile analysis of the results of femoropopliteal percutaneous transluminal angioplasty in the treatment of lower-extremity ischemia. *Journal of Vascular Surgery* **10**: 57–67.
Ades A. 2003. A chain of evidence with mixed comparisons: models for multi-parameter synthesis and consistency of evidence. *Statistics in Medicine*. **22**: 2995–3016.
Ades A, Sutton A. 2006. Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **169**: 5–35.
Agostino R. 1998. Tutorial in biostatistics propensity score methods for bias reduction in the comparison of a treatment to a non-randomised control group. *Statistics in Medicine* **2281**: 2265–2281.
Ahern J, Hubbard A, Galea S. 2009. Estimating the effects of potential public health interventions on population disease burden: a step-by-step illustration of causal inference methods. *American Journal of Epidemiology* **169**: 1140–1147.
Anonymous. 1992. Cross design synthesis: a new strategy for studying medical outcomes? *The Lancet* 944–946.
Austenfeld M, Thompson I, Middleton R. 1994. Meta-analysis of the literature: guideline development for prostate cancer treatment. *The Journal of Urology* **152**: 1866–1869.
Begg C. 1992. Book review: cross design synthesis: a new strategy for medical effectiveness research. *Statistics in Medicine* **11**: 1627–1630.
Begg C, Pilote L. 1991. A model for incorporating historical controls into a meta-analysis. *Biometrics* **47**: 899–906.
Benson K, Hartz A. 2000. A comparison of observational studies and randomized, controlled trials. *The New England Journal of Medicine* **342**: 1878–1886.
Bittl JA, He Y, Jacobs AK, Yancy CW, Normand S-LT. 2013. Bayesian methods affirm the use of percutaneous coronary intervention to improve survival in patients with unprotected left main coronary artery disease. *Circulation* **127**: 2177–2185.
Block P, Peterson E, Krone R, Kesler K, Hannan E, O'Connor G, Detre K. 1998. Identification of variables needed to risk adjust outcomes of coronary interventions: evidence-based guidelines for efficient data collection. *JACC* **32**: 275–82.
Brumback BA, Holmes LB, Ryan LM. 1999. Adverse effects of chorionic villus sampling: a meta-analysis. *Statistics in Medicine* **18**, 2163–2175.
Chelimsky E, Silberman G, Droitcour J. 1993. Cross design synthesis. *The Lancet* **341**, 498.
Clowse M, Behera M, Anders C, Copland S, Coffmann C, Leppert P, Bastian L. 2009. Ovarian preservation by GnRH agonists during chemotherapy: a meta-analysis. *Journal of Women's Health* **18**(3): 311–319.
Cole S, Stuart, E. 2010. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *American Journal of Epidemiology* **172**, 107–115.
Copas J. 2013. A likelihood-based sensitivity analysis for publication bias in meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **62**: 47–66.
Craxi A, Camma C, Giunta M. 1999. Definition of response to antiviral therapy in chronic hepatitis C. *Journal of Hepatology* **31**: 160–167.
Craxi A, DiBona D, Camma C. 2003. Interferon-alpha for HBeAg-positive chronic hepatitis B. *Journal of Hepatology* **39**, 99 – 105.
Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakrovitch C, Song F, Petticrew M, Altman DG. 2003. Evaluating non-randomised intervention studies. *Health Technology Assessment NHS R&D HTA Programme* **7**(27): 1–173.
Dias S, Welton N, Caldwell J, Ades A. 2010. Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine* **29**: 932–944.
Dias S, Sutton A, Ades A, Welton N. 2013. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making* **33**: 607–617.
Dominici F, Parmigiani G, Wolpert R, Hasselblad V. 1999. Meta-analysis of migraine headache treatments: combining information from heterogeneous designs. *Journal of the American Statistical Association* **94**: 16–28.
Donegan S, Williamson P, D'Alessandro U, Smith CT. 2012. Assessing the consistency assumption by exploring treatment by covariate interaction in mixed treatment comparison meta-analysis: individual patient level covariate versus aggregate trial level covariates. *Statistics in Medicine* **31**: 3840–3857.
Donegan S, Williamson P, D'Alessandro U, Garner P, Smith CT. 2013. Combining individual patient data and aggregate data in mixed treatment comparison meta-analysis: individual patient data may be beneficial if only for a subset of trials. *Statistics in Medicine* **32**: 914–930.
Droitcour J, Silberman G, Chelimsky E. 1993. A new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. *International Journal of Technology Assessment in Health Care* **9**: 440–9.
Eddy DM. 1989. The confidence profile method - a Bayesian method for assessing health technologies." *Operations Research* **37**: 210–28.
Eddy DM. 2013. "Top 10 projects. The Confidence Profile Method." *Available from*: http://www.davidmeddy.com/Top10projects.htm.
Eddy DM, Hasselblad V, Shachter R. 1990a. A Bayesian method for synthesizing evidence. The confidence profile method." *International Journal of Technology Assessment in Health Care* **6**: 31–55.

Eddy DM, Hasselblad V, Shachter R. 1990b. An introduction to a Bayesian method for meta-analysis: the confidence profile method. *Medical Decision Making* **10**: 15–23.
Eddy DM, Hasselblad V, Shachter R. 1992. Meta-analysis by the confidence profile method: the statistical synthesis of evidence. Academic Press, San Diego, CA.
Egger M, Ebrahim S, Davey Smith G. 2002. Where now for meta-analysis? *International Journal of Epidemiology* **31**: 1–5.
Evans J, Green J, Carlin P, Barrett L. 1995. Meta-analysis of antibiotic in tube thoracostomy. *The American Surgeon* **61**: 215–9.
Fortin M, Dionne J, Pinho G, Gignac J, Lapointe L. 2006. Randomized controlled trials: do they have external validity for patients with multiple comorbidities? *Annals Of Family Medicine* **4**: 104–108.
Frangakis C. 2009. The calibration of treatment effects from clinical trials to target populations. *Clinical Trials* **6**: 136–140.
Greenhouse J, Kaizar E, Kelleher K, Seltman H, Gardner W. 2008. Generalizing from clinical trial data: a case study. The risk of suicidality among pediatric antidepressant users. *Statistics in Medicine* **27**: 1801–1813.
Greenland S. 2005. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **168**: 267–306.
Grines C, Nelson T, Safian R, Hanzel G, Goldstein J, Dixon S. 2008. A Bayesian meta-analysis comparing AngioJet® thrombectomy to percutaneous coronary intervention alone in acute myocardial infarction. *Journal of Interventional Cardiology* **21**: 459–482.
Hall M, Chang S, Dalbagni G, Pruthi R, Seigne J, Skinner E, Wolf J, Schellhammer P. 2007. Guideline for the management of nonmuscle invasive bladder cancer (stages Ta, T1 and Tis): 2007 update. *The Journal of Urology* **178**: 2314–2330.
Hasselblad V. 1998. Meta-analysis of multi-treatment studies. *Medical Decision Making* **18**: 37–43.
Higgins J, Whitehead A. 1996. Borrowing strength from external trials in a meta-analyses. *Statistics in Medicine* **15**: 2733–2749.
Higgins J, Thompson S, Spiegelhalter D. 2009. A re-evaluation of random-effects meta-analysis. *Journal of Royal Statistical Society: Series A* **172**: 137–159.
Higgins JPT, Jackson D, Barrett JK, Lu G, Ades AE, White IR. 2012. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Research Synthesis Methods* **3**: 98–110.
Higgins J, Ramsay C, Reeves B, Deeks J, Shea B, Valentine J, Tugwell P, Wells G. 2013. Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions. *Research Synthesis Methods* **4**: 12–25.
Hinkelmann K, Kempthorne O. 1994. Design and analysis of experiments. Volume I: introduction to experimental design. John Wiley & Sons, New York.
Huang J, van Gelder J. 2002. The probability of sudden death from rupture of intracranial aneurysms: a meta-analysis. *Neurosurgery* **51**: 1001–1007.
Hurwitz E, Ake P, Adams A, Meeker W, Shekelle P. 1996. Manipulation and mobilization of the cervical spine. A systematic review of the literature. *Spine (Phila Pa 1976)* **21**(15): 1746–59.
Ibrahim J, Chen M. 2000. Power prior distributions for regression models. *Statistica science* **15**: 46–60.
Imai K, King G, Stuart E a. 2008. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **171**: 481–502.
Ioannidis J. 2010. Meta-research: the art of getting it wrong. *Research Synthesis Methods* **1**: 169–184.
Jackson C, Best N, Richardson S. 2006. Improving ecological inference using individual-level data. *Statistics in Medicine* **25**: 2136–2159.
Jackson C, Best N, Richardson S. 2008. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **171**: 159–178.
Jackson C, Best N, Richardson S. 2009. Bayesian graphical models for regression on multiple data sets with different variables. *Biostatistics* **10**: 335–351.
Jones W, Schmit K, Vemulapalli S, Subherwal S, Patel M, Hasselblad V, Heidenfelder B, Chobot M, Posey R, Wing L, Sanders G, Dolor R. 2013 May. Treatment strategies for patients with peripheral artery disease. *Rockville* (*MD*): *Agency for Healthcare Research and Quality* (*US*) Report No.: 13-EHC090-EF.
Kaizar E. 2011. Estimating treatment effect via simple cross design synthesis. *Statistics in Medicine* **30**: 2986–3009.
Klotzbuecher C, Ross P, Landsman P, Abbott T, Berger M. 2000. Patients with prior fractures have an increased risk of future fractures: a summary of the literature and statistical synthesis. *Journal of Bone and Mineral Research* **15**(4): 721–39.
Larose D, Dey D. 1997. Grouped random effects models for Bayesian meta-analysis. *Statistics in Medicine* **16**: 1817–1829.
Leach GE, Dmochowski RR, Appell RA. 1997. Female stress urinary incontinence clinical guidelines panel summary report on surgical management of female stress urinary incontinence. *Journal of Urology* **158**: 875–880.
Lefevre F, Aronson N. 2000. Ketogenic diet for the treatment of refractory epilepsy in children: a systematic review of efficacy. *Pediatrics* **105**(4): E46.

Lehmann H, Hinton R, Morello P, Santoli J. 2000. Developmental dysplasia of the hip practice guideline: technical report. *Pediatrics* **105**: 896–905.
Li Z, Begg C. 1994. Random effects models for combining results from controlled and uncontrolled studies in a meta-analysis. *Journal of the American Statistical Association* **89**: 1523–1527.
Licata A, DiBona D, Schepis F, Shahied L, Craxi A, Camma C. 2003. When and how to treat acute hepatitis C?" *Journal of Hepatology* **39**: 1056–1062.
Lu G, Ades A. 2004. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine* **23**. 3105–3124.
Lu G, Ades A. 2006. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* **101**(474): 447–459.
Lu G, Ades A. 2009. Modelling between-trial variance structure in mixed treatment comparisons. *Biostatistics* **10**: 792–805.
Lu G, Welton NJ, Higgins JPT, White IR, Ades AE. 2012. Linear inference for mixed treatment comparison meta-analysis: a two-stage approach. *Research Synthesis Methods* **2**(1): 43–60.
Lumley T. 2002. Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine* **21**(16): 2313–2324.
Lunn D, Spiegelhalter D, Thomas A, Best N. 2009. The BUGS project: evolution, critique and future directions. *Statistics in Medicine* **28**(25): 3049–3067.
Lunn D, Barrett J, Sweeting M, Thompson S. 2013. Fully Bayesian hierarchical modelling in two stages, with application to meta-analysis. *Journal of the Royal Statistical Society: Series C: Applied Statistics* **62**(4): 551–572.
Marshall E, Spiegelhalter D. 2007. Identifying outliers in Bayesian hierarchical models: a simulation-based approach. *Bayesian Analysis* **2**: 409–444.
McCarron C, Pullenayegum E, Thabane L, Goerree R, Tarride JE. 2010. The importance of adjusting for potential confounders in Bayesian hierarchical models synthesis evidence from randomized and non-randomized studies: an application comparing treatments for abdominal aortic aneurysms. *BMC Medical Research Methodology* **10**: 64.
McCarron C, Pullenayegum E, Thabane L, Goerree R, Tarride JE. 2011. Bayesian hierarchical models combining different study types and adjusting for covariate imbalances: a simulation study to assess model performance. *Plos One* **6**(10).
Mesgarpour B, Heidinger B, Schwameis M, Kienbacher C, Walsh C, Schmitz S, Herkner H. 2013. Safety of off-label erythropoiesis stimulating agents in critically ill patients: a meta-analysis. *Intensive Care Medicine* **39**(11): 1896–908.
Molitor J, Jackson C, Best N, Richardson S. 2009. Using Bayesian graphical models to model biases in observational studies and to combine multiple data sources: application to low birth weight and water. *Journal of the Royal Statistical Society: Series A* 32–50 **172**(3): 615–637.
Montague D, Barada JH, Belker AM. 1996. Clinical guidelines panel on erectile dysfunction: summary report on the treatment of organic erectile dysfunction. *Journal of Urology* **156**: 2007–2011.
Nanda K, Bastian L, Hasselbad V, Simel D. 1999. Hormone replacement therapy and the risk of colorectal cancer: a meta-analysis. *Obstetrics and Gynecology* **93**: 880–888.
Neuenschwander B, Branson M, Spiegelhalter D. 2009. A note on the power prior. *Statistics in Medicine* **28**: 3562–3566.
Nixon R, Duffy S. 2002. Cross-issue synthesis: potential application to breast cancer, tamoxifen and genetic susceptibility. *Journal of Cancer Epidemiology and Prevention* **7**: 205–212.
Norris S, Moher D, Reeves B, Shea B, Loke Y, Garner S, Anderson L, Tugwell P, Wells G. 2013. Issues relating to selective reporting when including non-randomized studies in systematic reviews on the effects of healthcare interventions. *Research Synthesis Methods* **4**: 36–47.
O'Rourke K, Walsh C. 2010. Impact of stroke units on mortality: a Bayesian analysis. *European Journal of Neurology* **17**: 247–251.
O'Rourke K, Walsh C, Hutchinson M. 2007. Outcome of beta-interferon treatment in relapsing-remitting multiple sclerosis: a Bayesian analysis. *Journal of Neurology* **254**: 1547–1554.
O'Rourke K, Walsh C, Kelly P. 2009. Safety and efficacy of IV-TPA for ischaemic stroke in clinical practice - a Bayesian analysis. *Cerebrovascular Diseases* **28**: 572–581.
Orr R. 1999. The impact of prophylactic axillary node dissection on breast cancer survival - a Bayesian meta-analysis. *Annals of Surgical Oncology* **6**(1): 109–116.
Committee on quality improvement. 2000. Clinical practice guideline: early detection of developmental dysplasia of the hip. *Pediatrics* **105**.
Peters J, Rushton L, Sutton A, Jones D, Abrams K, Mugglestone M. 2005. Bayesian methods for the cross-design synthesis of epidemiological and toxicological evidence. *Journal of the Royal Statistical Society, Series C* **54**: 159–172.
Peterson E, Shaw L, Califf R. 1997. Risk stratification after myocardial infarction. *Annals of Internal Medicine* **126**: 561–582.
Piepho H, Williams E, Madden L. 2012. The use of two-way linear mixed models in multitreatment meta-analysis. *Biometrics* **68**: 1269–1277.

Plummer M. 2003. JAGS: a program for analysis of Bayesian graphical models using gibbs sampling JAGS: Just Another Gibbs Sampler. *Proceedings of DSC*.

Prentice R, Langer R, Stefanick M, Howard B, Pettinger M, Anderson G, Barad D, Curb J, Kotchen J, Kuller L, Limacher M, Wactawski-Wende J. 2006. Combined analysis of Women's Health Initiative observational and clinical trial data on postmenopausal hormone treatment and cardiovascular disease. *American Journal of Epidemiology* **163**: 589–99.

Presanis AM, Ohlssen D, Spiegelhalter D, De Angelis D. 2013. Conflict diagnostic in directed acyclic graphs, with applications in Bayesian evidence synthesis. *Statistical Science* **28**(3): 376–397.

Prevost T, Abrams K, Jones D. 2000. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Statistics in Medicine* **19**: 3359–3376.

Reeves B, Higgins J, Ramsay C, Shea B, Tugwell P, Wells G. 2013. An introduction to methodological issues when including non-randomised studies in systematic reviews on the effects of interventions. *Research Synthesis Methods* **4**: 1–11.

Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, Boutitie F. 2008. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine* **27**: 1870–1893.

Rothstein HR, Sutton AJ, Borenstein M. 2005. Publication bias in meta-analysis: prevention, assessment and adjustment. Wiley, Chichester.

Salanti G, Higgins JP, Ades A, Ioannidis JP. 2008. Evaluation of networks of randomized trials. *Statistical Methods in Medical Research* **17**: 279–301.

Sampath S, Moran JL, Graham PL, Rockliff S, Bersten AD, Abrams KR. 2007. The efficacy of loop diuretics in acute renal failure: assessment using Bayesian evidence synthesis techniques. *Critical Care Medicine* **35**: 2516–2524.

Schmitz S, Adams R, Walsh C. 2013. Incorporating data from various trial designs into a mixed treatment comparison model. *Statistics in Medicine* **32**: 2935–2949.

Schünemann H, Tugwell P, Reeves B, Akl E, Santesso N, Spencer F, Shea B, Wells G, Helfand M. 2013. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. *Research Synthesis Methods* **4**: 49–62.

Segura J, Preminger G, Assimos D, Dretler S, Kahn R, Lingeman J, Macaluso J, McCullough D. 1994. Nephrolithiasis clinical guidelines panel summary report on the management of staghorn calculi. *The Journal of Urology* **151**: 1648–1651.

Segura J, Preminger G, Assimos D, Dretler S, Kahn R, Lingeman J, Macaluso J. 1997. Uteral stones clinical guidelines panel summary report on the management of ureteral calculi. *Journal of Urology* **158**: 1915–1921.

Senn S, Gavini F, Magrez D, Scheen A. 2011. Issues in performing a network meta-analysis. *Statistical Methods in Medical Research* **22**(2): 169–189.

Shachter RD, Eddy DM, Hasselblad V. 1990. An influence diagram approach to medical technology assessment. In Influence Diagrams, Belief Nets, and Decision Analysis, Oliver RM, Smith JQ (eds.). Wiley, Chichester; 321–350.

Shaw L, Peterson E, Kesler K, Hasselbad V, Califf R. 1996. A metaanalysis of predischarge risk stratification after acute myocardial infarction with stress electrocardiographic, myocardial perfusion, and ventricular function imaging. *American Journal of Cardiology* **78**: 1327–1337.

Shekelle P, Adams A, Chassin M, Hurwitz E, Brook R. 1992. Spinal manipulation for low-back pain. *Ann* **117**(7): 590–8.

Smith JA, Labasky RF, Cockett ATK, Fracchia JA, Montie JE, Rowland RG. 1999. Bladder cancer clinical guidelines panel summary report on the management of nonmuscle invasive bladder cancer (stages Ta, T1 and TIS). *The Journal of Urology* **162**: 1697–1701.

Soares MO, Dumville JC, Ades AE, Welton NJ. 2014. Treatment comparisons for decision making: facing the problems of sparse and few data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **177**: 259–279.

Spiegelhalter D, Best N. 2003. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in Medicine* **22**: 3687–3709.

Spiegelhalter D, Abrams KR, Myles JP. 2004. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England.

Sutton AJ, Abrams KR. 2001. Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research* **10**: 277–303.

Sutton A, Higgins J. 2008. Recent developments in meta-analysis. *Statistics in Medicine* **27**: 625–50.

Sutton A, Song F, Gilbody S, Abrams K. 2000. Modelling publication bias in meta-analysis: a review. *Statistical Methods in Medical Research* **9**: 421–445.

Sutton A, Kendrick D, Coupland C. 2008. Meta-analysis of individual- and aggregate-level data. *Statistics in Medicine* **27**: 651–669.

Thorlund K, Thabane L, Mills EJ. 2013. Modelling heterogeneity variance in multiple treatment comparison meta-analysis. Are informative priors the better solution? *BMC Medical Research Methodology* **13**: 2.

Turner R, Spiegelhalter D, Smith G, Thompson S. 2009. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **172**: 21–47.

Turner R, Lloyd J, Anumba D, Smith G, Spiegelhalter D, Squires H, Stevens J, Sweeting M, Urbaniak S, Webster R, Thompson S. 2012. Routine antenatal anti-dD prophylaxis in women who are RhD negative: meta-analyses adjusted for differences in study design and quality. *PLoS ONE 7* **2**: e30711.

Valentine J, Thompson S. 2013. Issues relating to confounding and meta-analysis when including non-randomized studies in systematic reviews on the effects of interventions. *Research Synthesis Methods* **4**: 26–35.

Verde PE. 2010. Meta-analysis of diagnostic test data: a bivariate Bayesian modeling approach." *Statistics in Medicine* **29**: 3088–3102.

Verde PE. 2013. Bamdit: Bayesian meta-analysis of diagnostic test data, r package version 1.1.

Wells GA, Shea B, Higgins J, Sterne J, Tugwell P, Reeves BC. 2013. Checklists of methodological issues for review authors to consider when including non-randomized studies in systematic reviews. *Research Synthesis Methods* **4**: 63–77.

Welton NJ, Ades AE, Carlin JB, Altman DG, Sterne JAC. 2009. Models for potentially biased evidence in meta-analysis using empirically based priors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **172**: 119–136.

White IR, Barrett JK, Jackson D, Higgins JPT. 2012. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. *Research Synthesis Methods* **3**: 111–125.

Wolpert RL, Mengersen KL. 2004. Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. *Statistical Science* **19**: 450–471.

Zimmerman M, Chelminski I, Posternak M. 2004. Exclusion criteria used in antidepressant efficacy trials: consistency across studies and representativeness of samples included. *The Journal of Nervous and Mental Disease* **192**: 87–94.