Statistics
in Medicine

Research Article

Received: 13 June 2016, Accepted: 16 December 2016 Published online 12 January 2017 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7223

Combining randomized and non-randomized evidence in network meta-analysis

Orestis Efthimiou,[a*†] Dimitris Mavridis,[a,b] Thomas P. A. Debray,[c,d] Myrto Samara,[e] Mark Belger,[f] George C. M. Siontis,[g] Stefan Leucht,[e] Georgia Salanti[a,h,i] and on behalf of GetReal Work Package 4

Non-randomized studies aim to reveal whether or not interventions are effective in real-life clinical practice, and there is a growing interest in including such evidence in the decision-making process. We evaluate existing methodologies and present new approaches to using non-randomized evidence in a network meta-analysis of randomized controlled trials (RCTs) when the aim is to assess relative treatment effects. We first discuss how to assess compatibility between the two types of evidence. We then present and compare an array of alternative methods that allow the inclusion of non-randomized studies in a network meta-analysis of RCTs: the naïve data synthesis, the design-adjusted synthesis, the use of non-randomized evidence as prior information and the use of three-level hierarchical models. We apply some of the methods in two previously published clinical examples comparing percutaneous interventions for the treatment of coronary in-stent restenosis and antipsychotics in patients with schizophrenia. We discuss in depth the advantages and limitations of each method, and we conclude that the inclusion of real-world evidence from non-randomized studies has the potential to corroborate findings from RCTs, increase precision and enhance the decision-making process. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: observational studies; observational evidence; observational data; multiple treatments meta-analysis; mixed treatment comparison; cohort studies

1. Introduction

Pairwise and network meta-analyses (NMAs) are often limited to synthesizing evidence from randomized controlled trials (RCTs). NMAs frequently disregard evidence from non-randomized studies (NRSs) because the authors assume that estimates of relative treatment effects are more likely to be biased, especially when confounding has been inadequately addressed. When non-randomized evidence is included in an NMA, this amplifies concerns about transitivity and consistency assumed by the method, and fears that results may be very precise, yet biased. But interest in including NRSs in the NMA synthesis and decision-making process is growing [1–5]. Although RCTs are considered the most reliable source of information on relative treatment effects, their strictly experimental setting and inclusion criteria may limit their ability to predict results in real-world clinical practice [6]. NRS-based estimates of treatment effects may complement evidence provided by RCTs and potentially address some of their

[a] *Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece*
[b] *Department of Primary Education, University of Ioannina, Ioannina, Greece*
[c] *Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands*
[d] *Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands*
[e] *Department of Psychiatry and Psychotherapy, Technische Universität München, München, Germany*
[f] *Eli Lilly and Company, Lilly Research Centre, Windlesham, U.K.*
[g] *Department of Cardiology, Bern University Hospital, Bern, Switzerland*
[h] *Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland*
[i] *Berner Institut für Hausarztmedizin (BIHAM), University of Bern, Bern, Switzerland*
**Correspondence to: Orestis Efthimiou, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece.*
[†] *E-mail: oremiou@gmail.com*

limitations. However, less than 4% of the NMAs published until the end of 2012 included at least one NRS (10 out of 261 identified NMAs) [7].

Expert opinion is required to formulate quantitative statements about the amount of bias propagated by NRSs in a body of evidence [8]. A recent review summarized methods that account for potential bias from non-randomized evidence in a pairwise meta-analysis [9]. These include an additive bias model that accounts for both external and internal biases in studies [10]; the confidence profile method [11]; likelihood adjustments [12]; and multiple-bias models [13]. Schmitz et al. [14] propose three different methods to combine data from different study designs in an NMA: ‘naïve’ pooling; use of non-randomized evidence as prior information; and a three-level hierarchical model.

In this paper, we present statistical methods for combining randomized and non-randomized evidence in an NMA, and we discuss their merits and limitations. We confine ourselves to the case where NMA is used to infer about the relative effects of health interventions. To this end, we consider only comparative NRSs that aim to estimate relative treatment effects. These include observational studies as well as comparative clinical trials that do not employ randomization.

2. Description of the motivating examples

2.1. *Percutaneous interventions for the treatment of coronary in-stent restenosis*

A previously published NMA synthesized aggregate data from 28 published RCTs (5914 patients) that compared eight different percutaneous interventional strategies for the treatment of coronary in-stent restenosis [15]. The follow-up in the included studies ranged from 6 to 60 months after the indexed intervention, and several clinical outcomes were considered; we focus on the dichotomous outcome ‘target–lesion revascularization’. Results were synthesized using the odds ratio (OR).

In addition to the RCTs, we identified data from six NRSs that provide evidence about target–lesion revascularization on five interventions. The studies included a total of 1113 patients in 14 different cohorts. The network is depicted in panel A, Figure 1. Detailed information about the included NRSs can be found in Section 4 of the Appendix.

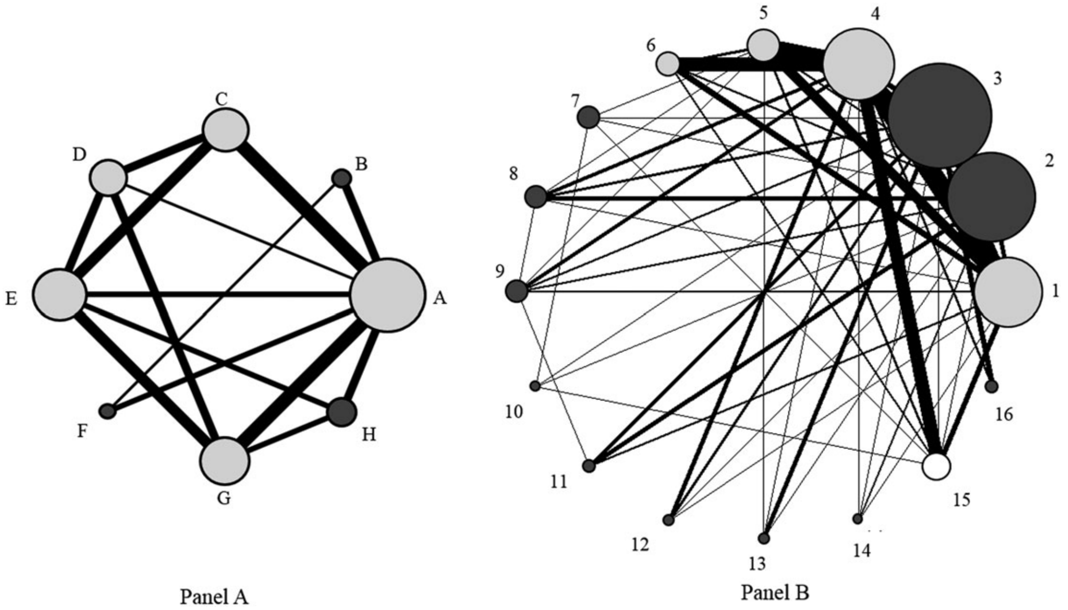

Figure 1. Networks of evidence for A) percutaneous interventions for coronary in-stent restenosis, and B) antipsychotics for schizophrenia. Dark grey nodes correspond to treatments compared in RCTs only; light grey nodes are examined in RCTs and NRSs. The size of each node is proportional to the number of studies that examine the corresponding treatment. The thickness of edges is proportional to the number of patients included in the studies that made the corresponding comparison. Codes for percutaneous interventions (Panel A): A = balloon angioplasty, B = bare metal stents, C = drug-coated balloons, D = everolimus-eluting stents, E = paclitaxel-eluting stents, F = rotablation, G = sirolimus-eluting stents, H = vascular brachytherapy.

2.2. *Antipsychotic drugs in schizophrenia*

The randomized evidence in this example consists of aggregate per-arm data from 167 RCTs (36 871 patients) which compared 15 antipsychotic drugs and placebo in patients diagnosed with schizophrenia [16]. Change in symptoms (efficacy) was measured 4–12 weeks after randomization, based on the brief psychiatric rating scale or the positive and negative syndrome scale. We use the standardized mean difference (SMD) to synthesize data. Using SMD as the effect measure enables the meta-analysis of studies that employ different scales, and it was also used in the original analysis [16]; researchers should note, however, that the use of SMD may be problematic under circumstances [17]. Study-level information was available for participant's mean age and duration of illness.

Non-randomized evidence consists of Individual Participant Data (IPD) from a large observational study (Schizophrenia Outpatient Health Outcome) with 8873 adult patients from 10 European countries, who were treated with five different antipsychotics during a 3-year time frame [18]. Short-term change in symptoms was measured at three months, based on the Clinical Global Impression scale. The network is depicted in panel B, Figure 1. Because we have signed non-disclosure agreements with our industry partner who provided the observational data we use in this example, we code treatments as 1–16.

3. Statistical methods

We present the methods assuming that relative treatments from non-randomized evidence have been estimated using valid epidemiological and statistical methods that aim to minimize bias if possible. An overview of such methods can be found in Faria et al. [3]. When considering the inclusion of NRS for which IPD are not available, extra caution is warranted as the aggregated reported effect estimates may originate from suboptimal analyses. In any case, the quality of the evidence provided by the identified NRSs needs to be critically appraised. The recently proposed ROBINS-I tool can be used to evaluate the risk of bias in estimates obtained from studies that did not use randomization [19]. If the identified NRSs are believed to have a very high risk of selection bias, their inclusion in NMA would be difficult to defend.

None of the NRSs about the effects of percutaneous interventions for coronary in-stent restenosis provided IPD. To estimate the SMDs for the antipsychotics from IPD in Schizophrenia Outpatient Health Outcome, we use regression adjustment analysis because there was enough overlap between the distributions of patient characteristics across treatment groups [3]. After consulting with expert psychiatrists to indicate important patient-level covariates, the SMDs from the non-randomized data were adjusted for baseline severity of the illness, age, gender and duration of illness.

In Section 3.1, we discuss the standard NMA model and fix notation. In Section 3.1, we describe methods for identifying possible discrepancies between randomized and non-randomized evidence. In Section 3.3, we present methods for synthesizing the two sources of evidence, assuming that no important differences have been found between them. We also describe the similarities and differences between the methods. Additional details for all models can be found in the Appendix.

3.1. *Standard NMA model with aggregated data*

The usual hierarchical, random-effects (RE) NMA model [20,21] synthesizes data from all available studies to estimate the summary treatment effect, μ_{XY}, and the heterogeneity standard deviation of the random effects, τ_{XY}, for each treatment comparison X vs. Y. Assume that for a two-arm XY study j, we observe the relative treatment effect d_{jXY}, along with a standard error, s_{jXY}. The model is then written as:

$$\begin{aligned} d_{jXY} &\sim N\left(\theta_{jXY}, s^2_{jXY}\right) \\ \theta_{jXY} &\sim N\left(\mu_{XY}, \tau^2_{XY}\right) \end{aligned} \tag{1}$$

We assume consistency, i.e. $\mu_{XY} = \mu_{XZ} - \mu_{YZ}$ for any choice of treatments X, Y and Z. This reduces the number of parameters and sets it equal to the number of treatments in the network minus one. Choosing treatment A as the reference, it is sufficient to estimate μ_{AX} for all treatments $X \neq A$; these are called the basic parameters. All other treatment effects can be obtained as linear combinations of the estimated μ_{AX}. Treatments can be ranked by any measure that summarizes the overlap between the estimated distributions of μ_{AX} [22]. An assumption commonly employed to facilitate the estimation of the heterogeneity parameters is to assume $\tau_{XY}=\tau$ for all treatments X and Y [23]. We use this assumption throughout this article for simplicity, although it is not necessary for any of the methods discussed below. For the inclusion of multi-arm studies, the model described above is expanded to allow for both within and between-study correlations of the observations by using multivariate distributions. For a comprehensive review of standard NMA methodology, we refer the reader to our recent publication [24].

3.2. Assessing the agreement within and across randomized and non-randomized evidence

An NMA of RCTs should be internally consistent: information from direct and indirect sources of evidence for each treatment comparison should agree. The presence of inconsistency can be tested statistically [25], and any disagreements could be explored in subgroup analysis or network meta-regression. In the presence of large unexplained inconsistency, NMA may be inappropriate. If no substantial inconsistencies are found in the NMA of RCTs, one can proceed with the evaluation of the agreement between randomized and non-randomized evidence. Disagreements between randomized and non-randomized evidence might be due to confounding in the non-randomized evidence, important differences in treatment effect modifiers across treatment groups of the NRSs, systematic differences in the definition of treatments between experimental and real-world settings (e.g. differences in dosages, modes of administration, adherence, etc.) or differences in patient characteristics. Inclusion criteria in RCTs are usually strict, so patients included in randomized studies may systematically differ from patients included in studies of real-world clinical settings.

Examination of differences between direct and indirect evidence can be formalized by graphical and statistical comparison of the following information for each pairwise comparison XY:

(1) direct randomized evidence, from XY randomized trials;
(2) indirect randomized evidence for XY, from the network, after excluding all direct XY studies;
(3) direct non-randomized evidence, from NRSs that include X and Y treatment arms.
(4) indirect, non-randomized evidence.

The four sources of evidence are independent, and they can be formally compared with statistical tests. The tests for inconsistency are, however, low powered, and the usefulness of formal statistical inference will be limited [26]. Note that differences between (i) and (ii) correspond to the notion of inconsistency in NMA [27].

Another, informal way to infer about agreement between randomized and NRSs is to compare the estimated heterogeneity parameters between the two different datasets. If NRSs are very different from RCTs, their inclusion into the network shall lead to an important increase in the heterogeneity parameter.

If researchers identify a source of disagreement between randomized and non-randomized evidence, they can perform analyses that account for it and improve comparability across the different sources of evidence. Cooper et al. [28] and Salanti et al. [29] presented a general network meta-regression framework for including study-level covariates in an NMA. We discuss network meta-regression models in more detail in Section 1 of the Appendix. Concerns about limited power and ecological bias are just as relevant in network meta-regression as in conventional meta-regression. Additionally, using network meta-regression in practice might prove to be difficult or even completely infeasible, e.g. if there is no usable information on important study-level covariates. Other models beyond network meta-regression might be useful in addressing specific sources of heterogeneity and inconsistency. Differences in the definitions of the treatments could be explicitly modeled using previously presented approaches [30,31]. When there are differences in the way that the outcomes are measured or reported (e.g. at different time points or using different scales), multiple outcomes NMA could be employed [32–34].

Researchers should note that even in the case of agreement between randomized and non-randomized evidence it may be inappropriate to perform a joint NMA. Epidemiological assessment of the compatibility of the various sources of data should always be performed by a content expert before undertaking any form of joint synthesis. The identified NRSs need to be examined on whether or not they are sufficiently similar in terms of population, intervention, comparator and outcomes with the ones specified in the research question that the review is aiming to answer [2]. When a NRS is deemed to be incompatible with the specific aims of the research question, it should be excluded from all analyses, irrespective of whether its findings happen to agree with those from the RCTs.

3.3. Methods for combining randomized and non-randomized evidence

If the NMA of RCTs is consistent and provided that there is no evidence of substantial disagreement between the randomized and the (adjusted) non-randomized evidence, synthesis of data is warranted. Below, we outline different approaches to synthesis and summarize their basic characteristics in Table I. We distinguish three main methodological approaches: *(i)* synthesis of non-randomized and randomized studies where information from NRSs is manipulated to reflect confidence in the study findings; *(ii)* constructing informative prior distribution from the non-randomized evidence and subsequently use it in the NMA of RCTs; and *(iii)* three-level hierarchical models, where one level of the model accounts for differences in RCT and NRS designs.

Before embarking on these adjusted analyses, investigators can obtain an initial insight of the effect of including NRSs in the network by synthesizing all data from all studies ('naïve' analysis) [14].

(A) *The 'Design-adjusted' evidence synthesis*

This approach synthesizes randomized and NRSs, after adjusting the mean effect sizes and/or the variance of the latter. In a two-armed NRS j the point estimate is shifted by β_j, where β_j is a bias term. The variance of the mean effect is also inflated (divided by a factor w_j, with $0 < w_j < 1$), so that the study's weight in the meta-analysis decreases. The investigator must define not only values for each set of β_j but also the direction of bias; that is which treatment is assumed to be 'favored'. The standard NMA model of Equation (1) is modified as follows:

$$d_{jXY} \sim N\left(\theta_{jXY} + \beta_j, \frac{s_{jXY}^2}{w_j}\right)$$
$$\theta_{jXY} \sim N\left(\mu_{XY}, \tau_{XY}^2\right)$$

For a multi-armed study j with T_j treatment groups, a bias vector $\boldsymbol{\beta_j}$ with $T_j - 1$ elements needs to be specified, along with w_j to inflate the within-study variance–covariance matrix of the observations. Estimation of heterogeneity variances can be performed as usual in NMA, e.g. using likelihood methods, the methods of moments or within a Bayesian framework. The parameters β_j or w_j can be set equal to fixed values or can be treated as random variables within a Bayesian framework. A Bayesian approach offers maximum flexibility and would allow us to assign prior distributions to β_j or w_j. Note that if a prior distribution is assigned to β_j then there is no need to do any additional down-weighting using w_j, because the additional uncertainty incorporated in β_j will result to a down-weighting of study j. These prior distributions can be specified to reflect potential deficiencies in the NRSs. For example, we may assign a prior for w_j centered to values above 0.5 or even close to one to a well-conducted NRS with low or moderate risk of bias and population, intervention, comparator and outcome characteristics that are relevant to the research question. In contrast, we could assign a distribution for w_j that is centered at low values below 0.5 or close to 0 to a NRS of high risk of bias or a NRS that does not quite address the research question. Alternatively, β_j or w_j can be treated as random variables exchangeable across the included NRSs [35]. In the presence of discrepancies between NRSs and RCTs and if w_j are treated as random variables, their posterior will be centered around smaller values limiting the effect of the NRSs [36].

Even though NRSs are generally considered to be at higher risk of bias [1], it may be hard to predict the direction (and also the magnitude) of possible biases in treatment effects [37–41], and this, in turn, makes it hard to pinpoint values for β_j and w_j. Biases in estimates of relative effects from NRSs may also depend on the method used to obtain them. Different methods used to estimate relative treatment effects from a NRS could give different results, making it harder to quantify potential biases. We recommend that researchers vary the amount of confidence they place on the non-randomized evidence by varying the value of w_j, and thus gauge the effect of non-randomized evidence on the model estimates. Setting all $w_j \approx 0$ corresponds to an NMA of only RCTs. Setting larger values w_j places more confidence in non-randomized estimates. Setting all $w_j = 1$ is equivalent to the naïve model described above, where results from the NRSs are taken at face value. We do not recommend using values $w_j > 1$, which place more confidence on the NRSs.

Eliciting expert opinion methods is needed to provide a range of plausible values for β_j and w_j. Turner et al. [8] discuss how to elicit expert opinion regarding the bias parameters, and how to pool information from different experts. We previously described another method that can be used to pool expert opinion

Table I. Overview of the presented approaches for combining randomized and non-randomized evidence.

	Design-adjusted analysis (Approach A)	Using informative priors (Approach B)	Three-level hierarchical models (Approach C)
Description	All trials are included in the NMA. Estimates from each NRS are adjusted for possible bias and over-precision.	Meta-analysis of RCTs using informative priors distributions formulated after meta-analyzing all NRSs.	Data is first synthesized by design, and then the design-specific summary estimates are pooled in a joint (network) meta-analysis.
How are NRS incorporated	Each NRS can be adjusted separately, according to its features. Alternatively, common bias parameters can be assumed for all NRS. Between-design variability in treatment effect is ignored.	The priors are shifted to account for bias, and/or the variances are inflated to down-weight estimates from NRSs. Between-design variability in treatment effect is ignored.	Each NRS can be adjusted separately, according to its features, if resources allow. Or adjustment for bias can be performed collectively for each design, on the design-level estimates.
Implementation challenges	Expert opinion is needed to choose appropriate values for w_j and β_j. Magnitude and directionality of bias in NRS may be hard to predict.	Choosing basic parameters and formulating priors may be non-trivial for complex network structures in the NRSs. Estimating heterogeneity may be hard if few RCTs or NRSs are available. Impossible to include NRSs and RCTs i n a joint network meta-regression.	Estimating τ_{des} either requires including several designs, or a strongly informative prior. Model C.1 requires meta-analyzing all designs, using a subset of the same basic parameters. Model C.2 is problematic in the presence of multi-arm studies
Software considerations	Easily implemented in all NMA software when fixed values for w_j and β_j are used.	Can be implemented only in a Bayesian framework (e.g. OpenBUGS)	Any software that implements hierarchical models
Better to use when	Should be preferred when resources allow inference about bias in each separate study.	Use when it is infeasible to infer about bias in each study separately.	Use when there are studies pertaining to multiple designs.
Technical details	The mean effect size in the j^{th} NRS can be shifted by a bias factor β_j. The variance of the treatment effects can be inflated by dividing with a variance-inflation factor $0 < w_j < 1$. When $w_j = 0$, all NRS are excluded; when $w_j = 1$, no adjustment takes place.	A RE-NMA of all NRSs is performed, and the predictive distributions for the summary effects are used as priors for the basic parameters of the NMA of RCTs.	First, data are meta-analyzed per design, using a design-specific heterogeneity parameter. Second, all design-specific estimates are synthesized to obtain an overall treatment effect that accounts for between-design heterogeneity.

NMA = network meta-analysis. RE = random-effects. NRS = non-randomized study. RCT = randomized controlled trial.

for a parameter ranging from 0 to 1 (such as w_j) which is based on beta distributions [32]. In this method each expert opinion can be weighted according to the expert's experience in the field. An approach alternative to expert opinion is to use information from previously published meta-epidemiological studies (e.g. Anglemyer et al. [41]) using a model such as the one presented by Welton et al. [42].

(B) *Using non-randomized evidence as prior information*

Most NMAs with RCTs are done within a Bayesian framework [7], and non-informative priors are typically assumed for all basic parameters μ_{AX}. In the presence of non-randomized evidence, however, these priors could be informative. The analysis is performed in two steps. First, a meta-analysis (or an NMA, if possible) uses the non-randomized evidence to estimate mean relative treatment effects for some (or all) of the basic parameters. The (posterior) distributions estimated in this first step can be 'adjusted' for bias (by adding a bias parameter β to the mean and dividing the variance by w). Then, these distributions are used in a second step as prior distributions for (some, or all of) the basic parameters of the NMA model which includes only RCTs. There are three different ways to construct informative priors. Expert opinion is needed for each one of them, for setting the values needed to adjust NRS.

The first approach is to start by synthesizing non-randomized evidence to estimate $\hat{\mu}_{AX}^{NRS}$ with a corresponding variance $\hat{V}_{AX}^{NRS}$. This estimated variance $\hat{V}_{AX}^{NRS}$ incorporates both the sampling error and heterogeneity, so that the distribution $N\left(\hat{\mu}_{AX}^{NRS},\ \hat{V}_{AX}^{NRS}\right)$ corresponds to the so-called 'predictive distribution' [43]. Then, down-weighting of the non-randomized evidence is achieved by assuming $\mu_{AX}{\sim}N\left(\hat{\mu}_{AX}^{NRS}+\beta_{AX},\ \hat{V}_{AX}^{NRS}/w_{AX}\right)$, where β_{AX} is a comparison-specific bias parameter and $0<w_{AX}<1$ is an inflation factor that quantifies the confidence to be placed in the non-randomized evidence for *AX* [14]. The model can be extended to incorporate uncertainty in these two parameters by assigning an informative prior distribution to β_{AX} or to w_{AX} (but not both). When the w parameters are assumed to be random, their posterior distribution depends on the agreement between the sources of evidence. When randomized and non-randomized evidence disagree, w will obtain smaller values and decrease the impact of the non-randomized evidence on the pooled estimates. Setting all $\beta_{AX}=0$ and $w_{AX}=1$ corresponds to placing full confidence in the non-randomized evidence; in such a case, and under the fixed-effect (FE) assumption, approach B becomes equivalent to a FE naïve analysis. Note that if some of the NRSs are multi-armed the corresponding estimates on basic parameters will be correlated. In such cases, we need to use multivariate prior distributions. For example, an A versus X versus Y NRS will provide information on μ_{AX}, μ_{AY} and the corresponding variance-covariance matrix; these can be used to formulate a bivariate normal prior distribution on the two basic parameters.

Another method for constructing the prior is to use the exact likelihood of the non-randomized data. The evidence provided by NRS j can be down-weighted by raising the likelihood to a power $0\leq a_j\leq 1$; this corresponds to the power-prior originally proposed by Ibrahim and Chen [44,45]. Dividing the variance by w_j (approach A) can be seen as a special case of the power prior method, where a study-specific power a_j is chosen for the likelihood of each study. In the special case of normally distributed data, approach A and the power prior approach are equivalent: a_j corresponds to w_j.

A third alternative approach is to use a mixture of priors [46,47]. The prior distribution consists of two parts; an informative prior formed by the predictive distribution as described above, and a flat (uninformative) prior. A factor $0\leq p\leq 1$ controls the amount of information drawn from the informative part of the prior and thus controls the contribution of non-randomized evidence to the final results. For the special case when normal distributions are used for both the informative and the uninformative parts of the prior, we can calculate how the mixture parameter p relates to the variance inflation parameter w. More details can be found in Section 3 of the Appendix.

The major difference between approaches A and B is the way that heterogeneity is accounted for in the analysis. In approach A, there is a single heterogeneity parameter for the relative treatment effects estimated in RCTs or NRSs. Approach B assumes two different heterogeneity parameters; one for NRSs and one for the RCTs. This is relevant because NRSs may tend to be more dissimilar than RCTs. When few NRSs are available (as in the example of antipsychotics), researchers cannot use a RE meta-analysis to formulate a predictive prior distribution, unless they make some assumption about heterogeneity. For example, the estimate from the naïve analysis or empirical data can use used to construct the predictive distributions. One important limitation of using NRSs to formulate prior distributions is that it precludes the option to explore differences with RCTs in a network meta regression model.

(C) *Three-level hierarchical models*

Three-level hierarchical models are suitable to synthesize data from studies with many different designs (e.g., different RCT designs, controlled trials, cohort studies, case–control, etc.) [14,48,49]. We present three different realizations of a hierarchical model for NMA. The main difference between them is the way the analysis incorporates the consistency assumption. We denote the various study designs with *i*. At the first level, each study is analyzed separately to obtain estimates of the relative effects of the treatments that are compared in the study.

Model C.1 At the second level in this model, NMA (including the consistency equations) is used to synthesize studies of the same design to obtain design-specific NMA estimates for the basic parameters and the heterogeneity variance τ_i^2.

$$d_{ijXY} \sim N\left(\theta_{ijXY}, s_{ijXY}^2\right)$$
$$\theta_{ijXY} \sim N\left(\kappa_{iAY} - \kappa_{iAX}, \tau_i^2\right)$$

The model can be used to include multi-arm studies using standard NMA methodology. Although we assume a single heterogeneity parameter within each design for simplicity, the model can be written using comparison-specific heterogeneities. At the third level, the basic parameters are assumed to be exchangeable across designs, which accounts for possible design-level heterogeneity, τ_{des}^2 , i.e. $\kappa_{iAX} \tilde{} N\left(\mu_{AX}, \tau_{des}^2\right), \kappa_{iAY} \tilde{} N\left(\mu_{AY}, \tau_{des}^2\right)$, etc. For the case of two different study designs (RCTs and NRSs), setting $\tau_{des}^2 = 0$ renders the model equivalent to approach B (with $w_{XA} = 1$ and $\beta_{AX} = 0$ for all basic parameters). Note that model C.1 requires an NMA to be performed for each study design using a subset of the same set of basic parameters. This might be infeasible in practice (see Section 4 of the Appendix for an illustration. Model C.1 cannot be used in such cases).

Model C.2 At the second level, we perform a series of pairwise, design-specific meta-analyses to obtain summary effects for all available treatment comparisons. The meta-analyses do not impose the consistency equations, but assume a common heterogeneity τ_i^2 for all treatment comparisons within design *i*, which corresponds to the 'unrelated mean effects model' described by Dias et al. [50]:

$$d_{ijXY} \sim N\left(\theta_{ijXY}, s_{ijXY}^2\right)$$
$$\theta_{ijXY} \sim N\left(\kappa_{iXY}, \tau_i^2\right)$$

If the common heterogeneity assumption is relaxed the model corresponds to a series of unrelated pairwise meta-analyses.

At the third level, a RE NMA (with heterogeneity τ_{des}^2) uses the consistency equations to synthesize estimates from the second level, i.e. we assume that $\kappa_{iXY} \tilde{} N\left(\mu_{AY} - \mu_{AX}, \tau_{des}^2\right)$. Model C.2 is problematic when the dataset contains multi-arm trials, because different parameterizations of the unrelated mean effects model can result into different estimates μ_{iXY} [50].

Model C.3 At the second level an NMA is performed separately per design. This imposes consistency exactly as in model C.1. Again the heterogeneity within each design can be assumed to be common, or different parameters can be assigned to different comparisons. At the third level, the estimates of the basic-parameters and their variance–covariance matrix from the design-level NMAs are synthesized in a new NMA, where they are treated as multi-arm studies and consistency equations are again imposed. This model assumes consistency twice—once within, and once across designs. At the third level, the common heterogeneity parameter of the NMA is τ_{des}^2. Note that this model does not require all NMAs at the second level to estimate a subset of the basic parameters.

In any three-level hierarchical model, the estimates from each study can be adjusted if appropriate, by shifting the mean and/or inflating the variance. Alternatively, evidence from each design can be downweighted by inflating the variance of the estimates obtained at the second level.

A major constraint in applying models C.1, C.2 and C3 is the availability of data from many different designs. Using an informative prior distribution for τ_{des}^2 will improve estimates in the presence of few designs. This prior distribution could be formulated using expert opinion that takes into account the expected dispersion of results in different designs. Alternatively, information of meta-epidemiological studies that include information on multiple types of designs (e.g. RCTs, registries, pragmatic trials, observational studies etc.) could be used to form empirical prior distributions.

3.4. Comparison of models

The three approaches presented here have similarities and, under specific circumstances, can lead to identical or equivalent statistical models. Frequentist and Bayesian implementation of model A with non-informative or weakly informative priors and the (necessarily Bayesian) Model B should give similar yet not identical results. Differences between frequentist and Bayesian methods are likely to result primarily from the estimation of heterogeneity parameters (which is assumed to be a random variable within the Bayesian context), as any prior distribution for these types of parameters tends to be informative. It is also possible to modify the models' characteristics, and to combine their distinctive features. For example, the effect sizes and variances of each NRS can be adjusted separately (as in model A) before using the three-level hierarchical model (C). Table I offers guidance for choosing a model appropriate to apply in practice.

4. Estimating the influence of the non-randomized evidence in the results from NMA

Once the non-randomized evidence is synthesized jointly with the randomized evidence, the credibility of the final summary treatment effects needs to be evaluated. The evaluation should consider, among other, the risk of bias in the included studies. One cannot exclude the possibility of residual bias due to the inclusion of NRSs, even after excluding studies with high risk of bias or adjusting effect estimates for relevant confounders. Although bias adjustments are usually considered only for the case of NRSs, RCTs may also suffer from design deficiencies that may bias their results. Calculating the relative contributions of the various sources of evidence is important when forming a critical appraisal of the NMA results. This is achieved by taking into account the various design limitations as well as how much each design is contributing to the final NMA treatment effect estimates [51].

It has been shown that each relative treatment effect estimated in an NMA can be expressed as the weighted average of a set of direct pairwise meta-analyses [52]. The weights of pairwise meta-analyses in estimating each NMA treatment effect form the 'hat matrix' [52] and can also be presented in a contribution matrix [51]. Because each direct meta-analysis is the weighted average of study-specific results, the contribution matrix can be updated to show the percentage contribution of each study in the pooled NMA estimates. A recent application of this approach can be found in [53]. The study contributions can be grouped to present how much evidence the randomized and NRSs finally contribute to the NMA results. The contribution matrix for an NMA can be obtained with available software, like the netmeta command in R [54] or the netweight command in Stata [55].

Contributions can be readily calculated in approach A when the model is fitted in a frequentist setting and can provide a percentage contribution of each study. A contribution matrix could also be estimated in approach C. In model C.1, it is possible to calculate the contribution each design makes to each meta-analysis of basic parameters at the third level, and then sum these contributions within each design. For models C.2 and C.3, the contribution matrix of the NMA, which is performed at the third level (the design level), contains information on the effect of each design on the pooled summary effects.

Calculating study contributions in approach B (and generally for models fitted in Bayesian setting) is not straightforward. A measure similar to a multivariate I^2 statistic introduced by Jackson et al. can be used to estimate contributions (I_R^2 in [56]). In an NMA model, the $(T-1)^{th}$ root of the determinant of the variance-covariance matrix (denoted here as γ) measures the precision of the estimates for the basic parameters. The contribution of the non-randomized evidence can then be approximated by the relative decrease in γ when priors with various weights are employed. For example, the model is fitted using uninformative prior, so that γ_{RCT} is based only on RCTs, and then γ_{all} is estimated using informative priors based on NRSs. Then, the contribution of NRS can be approximated with $I_{NRS}^2 = (\gamma_{RCT} - \gamma_{all})/\gamma_{RCT}$. I_{NRS}^2 can also be used in models A and C (or any other similar model), but complications arise because the estimate for heterogeneity may change after NRSs are included and the I_{NRS}^2 may acquire negative values for these models. Also, it might be resource intensive to use it to calculate the contribution of each individual study.

5. Illustrative examples

We used the network of the percutaneous interventions for the treatment of coronary in-stent restenosis to illustrate approach A, and we employed the network command in Stata [57,58]. The network of antipsychotics for schizophrenia was used to illustrate approach B. The mean age and mean duration of illness of participants are important study-level effect modifiers, and we used network meta-regression to include them in the analysis. Because estimating heterogeneity (τ) in the non-randomized evidence is impossible in this example (there is only one NRS), we used the estimate from the naïve analysis to construct predictive distributions. To exemplify approach C, we used the antipsychotics network, after assuming four different study categories; RCTs at low, moderate and high risk of bias (see Section 6 of the Appendix for more detail), and also the observational study. We then treated the groups as different designs. We used model C1 presented in Section 3.3 to synthesize the data. We fitted approaches B and C in OpenBUGS version 3.2.2 [59,60] (codes are provided in the Appendix). For all analyses, we ran two chains that allowed for 200 000 samples after a 100 000 burn-in period, and checked convergence with the Gelman and Rubin diagnostic. Initial values were automatically generated using the default OpenBUGS procedure.

In order to assess compatibility between the various sources of evidence we first decomposed the randomized evidence into its direct and indirect components, and then compared them with the non-randomized evidence. We used the network command in Stata, which implements the loop-specific approach [61] and the node-splitting approach [27].

5.1. Assessing the agreement within and across randomized and non-randomized evidence

5.1.1. Percutaneous interventions for coronary in-stent restenosis. In Figure 2, we present the relative treatment effects estimated from direct, indirect randomized and non-randomized evidence. For each treatment comparison, the figure shows the treatment effects from the direct randomized evidence (from RCTs that compare the corresponding treatments), the indirect randomized evidence (from the network of RCTs after excluding direct evidence), the direct non-randomized evidence (from NRSs that compare the corresponding treatments) and the indirect non-randomized evidence (from the network of NRSs after excluding direct non-randomized evidence). Randomized and non-randomized evidence are in reasonable agreement for most treatment comparisons. However, for some comparisons (in particular DvsA, DvsC, EvsC and GvsB), there are considerable discrepancies. The discrepancies in EvsC and DvsC are mainly driven by a single EC NRS [62] whose results were very different from those in RCTs examining the same comparison. This might be due to chance or indicate important unaccounted confounding in the NRS and differences in the population, setting and interventions between the NRS and the RCTs.

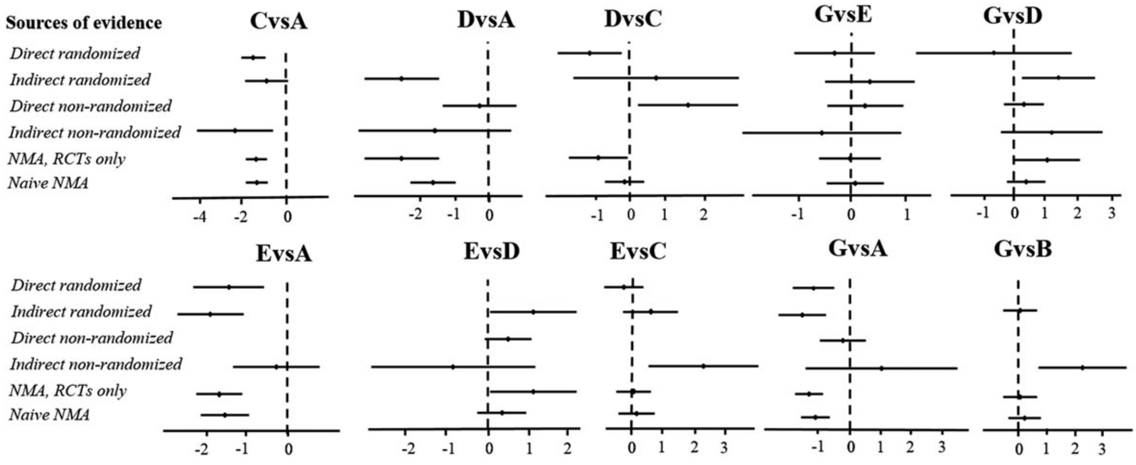

Figure 2. Relative treatment effects (log OR and their 95% confidence intervals) for target–lesion revascularization with percutaneous interventions for coronary in-stent restenosis, estimated from the various sources of evidence. The results from NMA using only RCTs and from NMA using both RCTs and NRS (naïve NMA) are also presented. Only treatment comparisons informed by both randomized and non-randomized evidence are presented. Codes of interventions as per Figure 1.

The heterogeneity standard deviation was estimated to be $\hat{\tau} = 0.36$ for the RCTs network and $\hat{\tau} = 0.45$ for the naïve analysis. The loop-specific approach did not detect any inconsistencies in the RCTs network or in the combined RCTs and NRSs network. The node-splitting approach showed no inconsistencies in the RCTs network, but found inconsistency in the DvsA comparison in the combined RCTs and NRSs NMA. This inconsistency is due to the difference in estimates coming from indirect randomized and direct non-randomized evidence as shown in Figure 2. Detailed results from these analyses are presented in Section 5 of the Appendix.

In a real-life application, researchers would need to further explore these disagreements, scrutinizing the identified studies in order to make a judgment on whether or not it is appropriate to pool them in a single NMA. To further exemplify the methods, we assume that the identified differences are due to chance, and we proceed into synthesizing the totality of the evidence in an NMA.

5.1.2. Antipsychotic drugs in schizophrenia. Figure 3 presents the estimates for treatment comparisons for which both randomized and non-randomized evidence is available. We detected some disagreement only for the 4v6 comparison. The confidence intervals of effects from direct randomized and non-randomized evidence overlap to some extent, but there is a discrepancy between the estimates that correspond to indirect randomized evidence and the evidence from the single NRS. This might indicate that the adjustment of the non-randomized data was insufficient (e.g., due to residual confounding). There is considerable heterogeneity in the five RCTs that make up the direct randomized evidence ($I^2 = 71\,\%, \hat{\tau} = 0.33$), while the prediction interval is rather broad (-1.18, 1.23) and includes the estimate from the NRS. In such a case, heterogeneity in the RCTs that compare 4v6 should be fully explored.

Comparing estimates from an NMA of the RCTs only with estimates from the naïve NMA also indicated that treatment effects are in agreement between the different sources. The precision of some treatment comparisons increased when non-randomized evidence were included in the network. For instance, the width of the credible interval was reduced by 27% for the 1vs15 comparison. We compared the heterogeneity standard deviation before and after including the NRS and found no differences: $\hat{\tau} = 0.08$ (95% Cr.I. (0.04, 0.12)) for the RCTs-only network, and 0.09 (0.05, 0.12) for the naïve RCTs and NRS network. The loop-specific and the node-splitting approach did not provide any strong evidence of inconsistency either before or after adding the NRS to the network of RCTs (results are presented in Section 7.1 of the Appendix). We conclude that we found no evidence of important disagreement between randomized and non-randomized summary effects and that synthesis of both sources is warranted.

5.2. *Using the design-adjusted analysis (approach A) for the coronary in-stent restenosis example*

We employ approach A where we assume that estimates from each NRS are expected to be unbiased ($\beta = 0$), with an uncertainty reflected on w. However, we want to down-weight the NRS employing various scenarios for a common variance-inflation factor (w), for all NRSs.

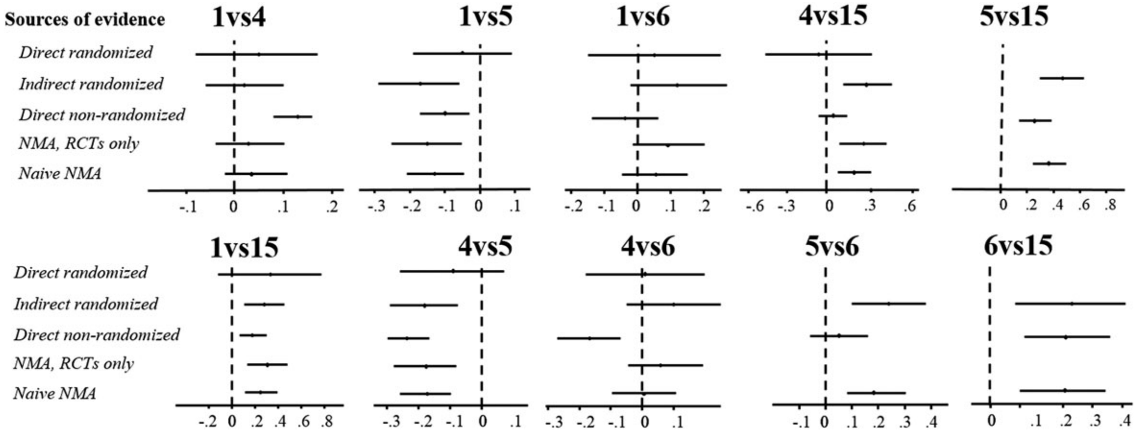

Figure 3. Relative treatment effects (standardized mean differences SMD and their 95% credible intervals) for improvement in symptoms scale with antipsychotics in patients with schizophrenia estimated from the various sources of evidence. The results from NMA using only RCTs and from NMA using both RCTs and NRS (naïve NMA) are also presented. Only treatment comparisons informed by both randomized and non-randomized evidence are presented.

Figure 4 shows the estimated relative treatment effects for all comparisons informed by both randomized and NRSs for various values of confidence placed on NRSs. For some treatment comparisons (CvsA; DvsA; EvsA; GvsA), the inclusion of the NRSs in the network corroborated the findings of the NMA based on RCTs alone and increased the precision in the estimates. For three comparisons (DvsC; EvsD; GvsD), the inclusion of non-randomized evidence pulled the summary effect towards the non-effect line even for low values of w. Such a result might potentially shed doubts on whether or not the difference in the efficacy of the corresponding interventions can be translated into difference of real-world effectiveness, and might have interesting clinical implications. However, these changes in the summary effect and their confidence intervals were not sufficient to change the treatment hierarchy estimated using the surface under the cumulative ranking line [63] which remained unchanged with the various analysis models. Results can be found in Section 5 of the Appendix.

In this example, the NRSs accounted for 16% of the total sample size in the network. In the naïve analysis, the contribution of NRSs was 20.3%. The contribution decreased to 19.0% for $w = 0.8$; 16.5% for $w = 0.5$ and 10.3% for $w = 0.2$.

5.3. *Using non-randomized evidence as prior information (approach B) for the schizophrenia example*

In order to use the non-randomized evidence as prior information, we first needed to choose the basic parameters appropriately (see Section 2 of the Appendix on choosing the basic parameters). The single NRS compares treatments 1, 4, 5, 6 and 15, and consequently we could choose any of those treatments to be the reference treatment; here, we chose treatment 1. We used the NRS to estimate a multivariate predictive distribution for the basic parameters 1vs4, 1vs5, 1vs6 and 1vs15, and we used the heterogeneity estimated from the naïve analysis ($\hat{\tau} = 0.09$). We used a common variance inflation factor for these

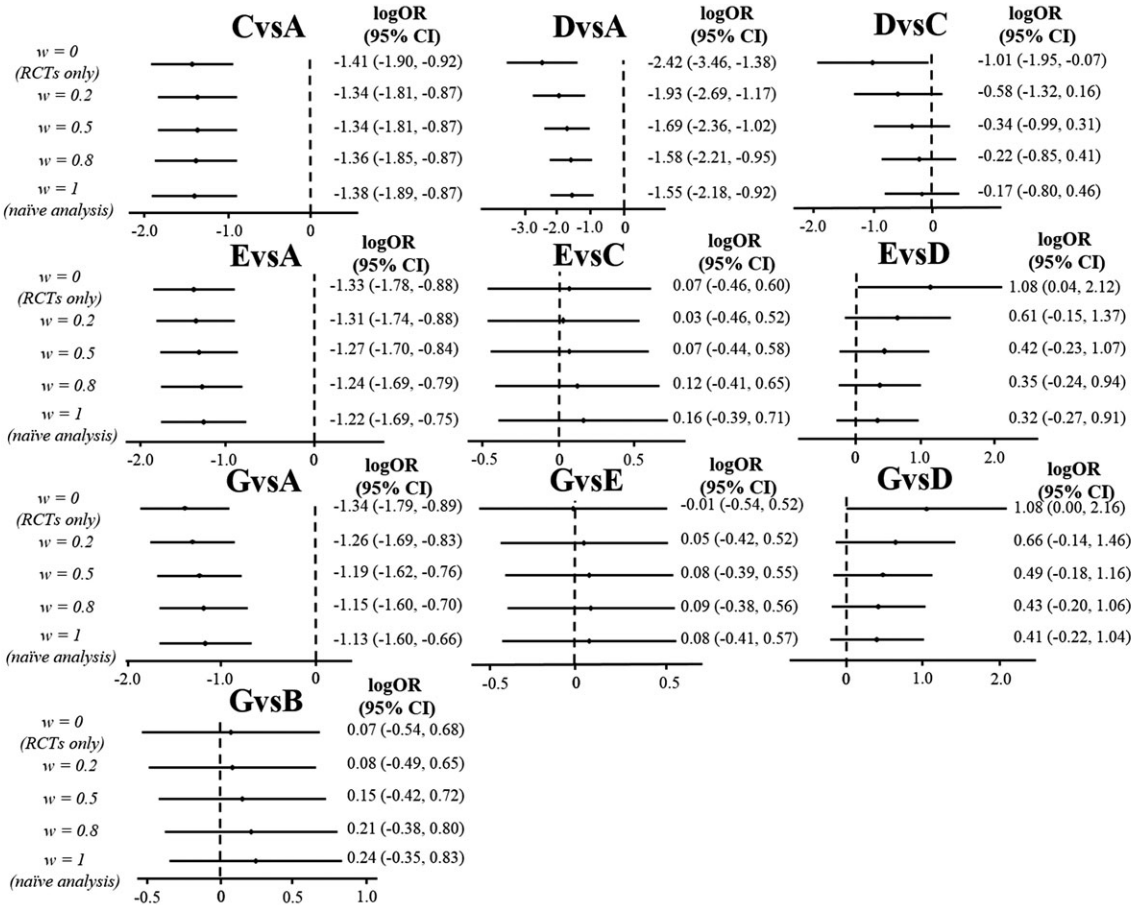

Figure 4. Relative treatment effects (log OR and their 95% confidence intervals) for target–lesion revascularization with percutaneous interventions for coronary in-stent restenosis estimated using approach A. Evidence from the NRSs is given increasing weight with the parameter w. Only treatment comparisons informed by both randomized and non-randomized evidence are presented. Codes of interventions as per Figure 1.

basic parameters to explore three different scenarios: $w \sim Unif(0, 0.3)$, which places a low level of confidence in the non-randomized evidence; $w \sim Unif(0.3, 0.7)$, which places a medium level of confidence; and $w \sim Unif(0.7, 1)$, which places a high level. The rest of the basic parameters were drawn from non-informative distributions, $N(0, 100^2)$.

Estimates of the relative treatment effects are presented in Figure 5. For most treatment comparisons (1vs5, 1vs15, 4vs5, 4vs15, 5vs6, 5vs15 and 6vs15), the inclusion of the single NRS in the evidence-base confirmed the findings and increased precision of the estimates. As expected, treatment hierarchy was robust to the various prior options. Estimates of the model parameters and surface under the cumulative ranking values are presented in Section 7 of the Appendix.

For the first scenario, I^2_{NRS} was calculated to be 4.7%; for the second scenario, it was 7%; and, for the third, it was 8.5%.

5.4. *Using the three-level hierarchical model (model C1) for the schizophrenia example*

We assumed that the design 'high RoB RCTs' and the 'NRS' design have the same heterogeneity parameter denoted as τ_{high}, and that designs 'low RoB' and 'moderate RoB' pertain to $\tau_{low\,\&moderate}$.

We used four different values for w ranging from 0 (NRS excluded from the analysis) up to 1 (no down-weighting of the NRS). In Table II, we present the model estimates only for the basic parameters which were informed by both randomized and non-randomized sources. Comparing to the results of approach B, the most notable difference is the increased imprecision. This is because model C incorporates an additional source of variability in the model, i.e. the heterogeneity across designs. Another interesting finding is that NRS had a larger impact on results; for no down-weighting ($w = 1$), I^2_{NRS} was calculated to be 14% (for the third scenario of model B in the previous section this was 8.5%). This increase of the contribution of NRS was because this study was the only source of information for one of the four available designs in the dataset.

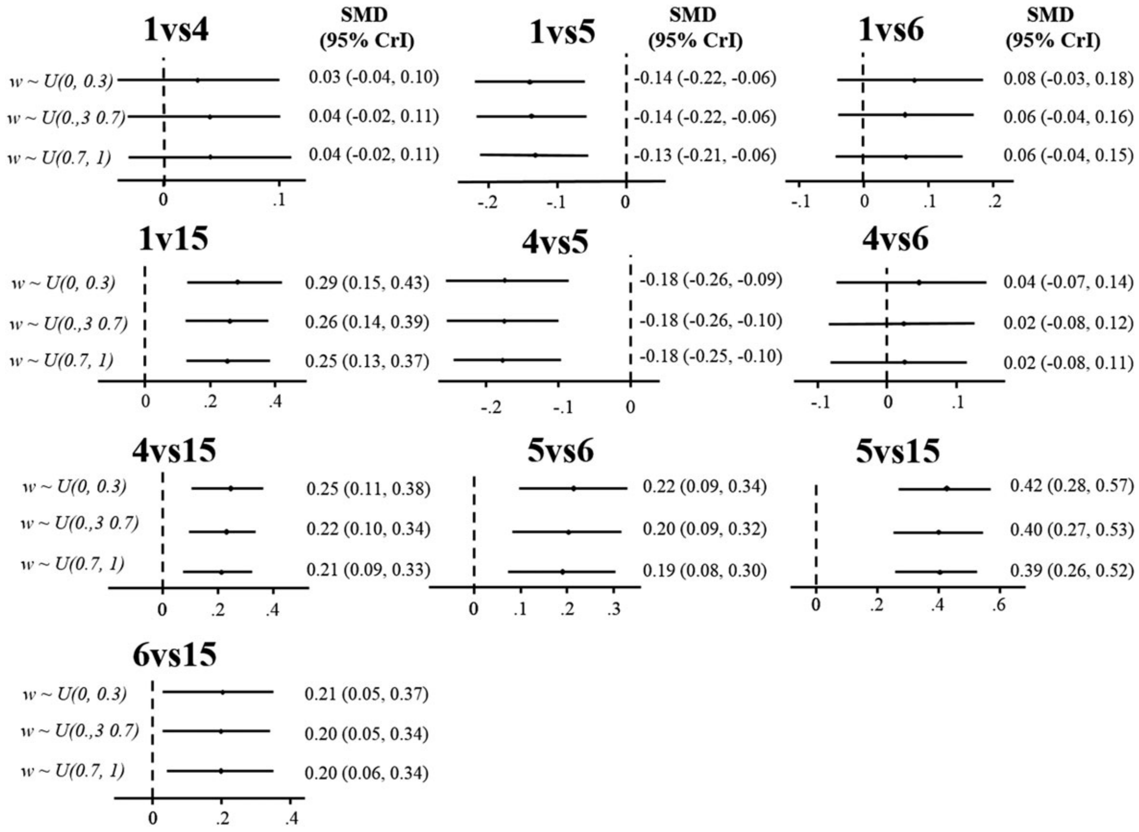

Figure 5. Relative treatment effects (standardized mean differences SMD and their 95% credible intervals) for improvement in symptoms scale with antipsychotics in patients with schizophrenia. A large NRS is contributing data to form an informative prior (approach B). Evidence from the NRS is given increasing weight with the parameter *w*. Only treatment comparisons informed by both randomized and non-randomized evidence are presented.

Table II. Estimated effects and 95% credible intervals for the antipsychotics network, for basic parameters informed both by RCTs and NRS as estimated from model C1. Larger values of *w* correspond to larger down-weighting of the NRS.

	Treatment comparison			
Variance inflation factor *(w)*	1v4	1v5	1v6	1v15
0.0 (RCTs only)	0.04 (−0.08, 0.16)	−0.12 (−0.26, 0.01)	0.08 (−0.08, 0.24)	0.31 (0.11, 0.50)
0.3	0.05 (−0.04, 0.16)	−0.12 (−0.23, −0.01)	0.05 (−0.08, 0.18)	0.26 (0.10, 0.41)
0.7	0.06 (−0.04, 0.16)	−0.11 (−0.22, −0.01)	0.04 (−0.08, 0.17)	0.25 (0.10, 0.40)
1.0 (no down−weighting)	0.06 (−0.04, 0.16)	−0.11 (−0.22, 0.00)	0.04 (−0.08, 0.17)	0.25 (0.11, 0.41)

For $w=1$, the heterogeneity standard deviations were estimated $\hat{\tau}_{high}=0.07$ (0.01, 0.12) and $\hat{\tau}_{low\&moderate}=0.06$ (0.003, 0.16). The estimate for the design-level heterogeneity was $\hat{\tau}_{des}=0.07$ (0.02, 0.13), indicating relatively small differences across different designs. These estimates did not materially change when we used different weighting schemes for the NRS.

6. Discussion

In this article, we discussed approaches for incorporating non-randomized evidence in an NMA of RCTs. These approaches should be employed after researchers have performed a formal assessment of the risk of bias and the applicability of the identified studies. This is needed in order to evaluate whether the inclusion of each study is sufficiently justified. We also argue that before performing a joint analysis of randomized and non-randomized evidence, researchers need to ensure the compatibility of the different pieces of evidence, for each treatment comparison. If studies are deemed incompatible a priori (i.e. before comparing effect estimates across study designs), their inclusion in the NMA should not be considered.

We grouped the available methods for combining randomized and non-randomized evidence in an NMA into three categories; the design-adjusted analysis, the use of informative prior distributions and the three-level hierarchical models. We do not recommend using the naïve approach as the main method of analysis, but it can be a useful starting point, and can provide insight about the effect of including non-randomized evidence in the analysis. The naïve approach can also be used to assess compatibility of randomized and non-randomized evidence, via monitoring changes in network heterogeneity and inconsistency before and after inclusion of non-randomized evidence.

The ‘design-adjusted’ approach extends the naïve approach by considering the design of the studies. The data from NRSs are ‘shifted’ and down-weighted based on external opinion about their credibility. We recommend using this approach when resources allow for a separate assessment of bias for each NRS. Using the non-randomized evidence to construct informative prior distributions for the basic parameters of the model is an elegant alternative for including non-randomized evidence in the NMA. A key difference with the design-adjusted approach is in the estimation of heterogeneity, which is performed separately for RCTs and NRSs. This approach might be more intuitive for clinicians, because they typically have prior opinions about treatment effects based on their experience with patient follow-up, monitoring and registries. Hierarchical models are more appropriate when data from studies of several different designs are to be synthesized and account for heterogeneity within and across designs. While the other methods assume that the underlying treatment effects are the same across designs, the three-level hierarchical models assume that the treatment effects are different—but exchangeable—across different types of studies.

In our two illustrative examples, we employed the design-adjusted approach (in-stent restenosis) and the prior-based approach (schizophrenia), with various degrees of confidence placed to the NRSs. For the in-stent restenosis network, the inclusion of NRSs confirmed the findings of the RCTs-only analysis for most comparisons. For some of the comparisons, results were shifted, indicating smaller differences in the outcome between the interventions even when low confidence was placed on the non-randomized evidence. For the schizophrenia example, the inclusion of non-randomized evidence did not materially impact on the conclusions of the analysis. Precision of the relative treatment effect estimates increased only slightly when we incorporated non-randomized evidence in the analysis, because the contribution of the single, although very large, NRS was small compared to that of 167 RCTs.

Whatever method researchers choose to employ, they should keep in mind that it is difficult to predict the magnitude or direction of possible biases introduced by including NRSs in an NMA. We thus advise them to explore the effect of placing different levels of confidence in the non-randomized evidence before they draw final conclusions. We also recommend that all results should be evaluated after considering the relative contribution of each source of evidence in the pooled estimates. This might be especially relevant for the case that NRSs are used to connect disconnected parts of the network of RCTs; on such occasions, the connecting studies may acquire an unduly large contribution for (some) of the network estimates, even after severe down-weighting.

Funding

This work was supported by the Innovative Medicines Initiative Joint Undertaking under grant agreement No 115546, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007–2013) and EFPIA companies' in kind contribution. The research leading to these results was conducted as part of the GetReal consortium. For further information, please refer to www.imi-getreal.eu. This paper only reflects the personal views of the stated authors.

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

Acknowledgements

We would like to thank the two anonymous reviewers for their constructive criticism and insightful comments, which helped us largely improve our paper.

References

1. Reeves BC, Higgins JPT, Ramsay C, Shea B, Tugwell P, Wells GA. An introduction to methodological issues when including non-randomised studies in systematic reviews on the effects of interventions. *Research Synthesis Methods* 2013; **4**:1–11.
2. Schünemann HJ, Tugwell P, Reeves BC, Akl EA, Santesso N, Spencer FA, Shea B, Wells G, Helfand M. 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* 2013; **4**:49–62.
3. Faria R, Hernadez Alava M, Manca A, Wailoo AJ. The use of observational data to inform estimates of treatment effectiveness in technology appraisal: methods for comparative individual patient data. *NICE DSU technical support document 17*, 2015. http://www.nicedsu.org.uk/Observational-data-TSD*(2973296).htm*.
4. Cameron C, Fireman B, Hutton B, Clifford T, Coyle D, Wells G, Dormuth CR, Platt R, Toh S. Network meta-analysis incorporating randomized controlled trials and non-randomized comparative cohort studies for assessing the safety and effectiveness of medical treatments: challenges and opportunities. *Systematic Reviews* 2015; **4**:147.
5. Bell H, Wailoo AJ, Hernandez M, Grieve R, Faria R, Gibson L, Grimm S: The use of real world data for the estimation of treatment effects in NICE decision making. *NICE DSU technical support document*, 2016. http://www.nicedsu.org.uk/Real-World-Data-RWD*(3026863).htm*.
6. Rothwell PM. External validity of randomised controlled trials: "To whom do the results of this trial apply?". *The Lancet* 2005; **365**:82–93.
7. Nikolakopoulou A, Chaimani A, Veroniki AA, Vasiliadis HS, Schmid CH, Salanti G. Characteristics of networks of interventions: a description of a database of 186 published networks. *PLoS ONE* 2014; **9**:e86754.
8. Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2009; **172**:21–47.
9. Verde PE, Ohmann C. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. *Research Synthesis Methods* 2015; **6**:45–62.
10. Spiegelhalter DJ, Best NG. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in Medicine* 2003; **22**:3687–3709.
11. Eddy DM, Shachter R. *Meta-Analysis by the Confidence Profile Method* (Har/Dskt edn). Academic Press: Boston, 1992.
12. Wolpert RL, Mengersen KL. Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. *Statistical Science* 2004; **19**:450–471.
13. Greenland S. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2005; **168**:267–306.
14. Schmitz S, Adams R, Walsh C. Incorporating data from various trial designs into a mixed treatment comparison model. *Statistics in Medicine* 2013; **32**:2935–2949.
15. Siontis GCM, Stefanini GG, Mavridis D, Siontis KC, Alfonso F, Pérez-Vizcayno MJ, Byrne RA, Kastrati A, Meier B, Salanti G, Jüni P, Windecker S. Percutaneous coronary interventional strategies for treatment of in-stent restenosis: a network meta-analysis. *The Lancet* 2015; **386**:655–664.

16. Leucht S, Cipriani A, Spineli L, Mavridis D, Orey D, Richter F, Samara M, Barbui C, Engel RR, Geddes JR, Kissling W, Stapf MP, Lässig B, Salanti G, Davis JM. Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: a multiple-treatments meta-analysis. *Lancet* 2013; **382**:951–962.
17. Deeks JJ, Higgins JP, Altman DG. Analysing Data and Undertaking Meta-Analyses. In *Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series*, Higgins JP, Green S (eds). John Wiley & Sons, Ltd: Chichester, UK, 2008.
18. Haro JM, Edgell ET, Jones PB, Alonso J, Gavart S, Gregor KJ, Wright P, Knapp M. SOHO Study Group: The European Schizophrenia Outpatient Health Outcomes (SOHO) study: rationale, methods and recruitment. *Acta Psychiatrica Scandinavica* 2003; **107**:222–232.
19. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR, Chan AW, Churchill R, Deeks JJ, Hróbjartsson A, Kirkham J, Jüni P, Loke YK, Pigott TD, Ramsay CR, Regidor D, Rothstein HR, Sandhu L, Santaguida PL, Schünemann HJ, Shea B, Shrier I, Tugwell P, Turner L, Valentine JC, Waddington H, Waters E, Wells GA, Whiting PF, Higgins JP. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016; **355**:i4919.
20. Salanti G, Higgins JP, Ades AE, Ioannidis JP. Evaluation of networks of randomized trials. *Statistical Methods in Medical Research* 2008; **17**:279–301.
21. Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making* 2013; **33**:607–617.
22. Nikolakopoulou A, Mavridis D, Salanti G. Planning future studies based on the precision of network meta-analysis results. *Statistics in Medicine* 2016; **35**:978–1000.
23. Higgins JPT, Whitehead A. Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine* 1996; **15**:2733–2749.
24. Efthimiou O, Debray TPA, van Valkenhoef G, Trelle S, Panayidou K, Moons KGM, Reitsma JB, Shang A, Salanti G. on behalf of GetReal Methods Review Group: GetReal in network meta-analysis: a review of the methodology. *Research Synthesis Methods* 2016; **7**:236–263.
25. Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* 2006; **101**:447–459.
26. Veroniki AA, Mavridis D, Higgins JP, Salanti G. Characteristics of a loop of evidence that affect detection and estimation of inconsistency: a simulation study. *BMC Medical Research Methodology* 2014; **14**:106.
27. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine* 2010; **29**:932–944.
28. Cooper NJ, Sutton AJ, Morris D, Ades AE, Welton NJ. Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. *Statistics in Medicine* 2009; **28**:1861–1881.
29. Salanti G, Marinho V, Higgins JPT. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. *Journal of Clinical Epidemiology* 2009; **62**:857–864.
30. Del Giovane C, Vacchi L, Mavridis D, Filippini G, Salanti G. Network meta-analysis models to account for variability in treatment definitions: application to dose effects. *Statistics in Medicine* 2013; **32**:25–39.
31. Warren FC, Abrams KR, Sutton AJ. Hierarchical network meta-analysis models to address sparsity of events and differing treatment classifications with regard to adverse outcomes. *Statistics in Medicine* 2014; **33**:2449–2466.
32. Efthimiou O, Mavridis D, Cipriani A, Leucht S, Bagos P, Salanti G. An approach for modelling multiple correlated outcomes in a network of interventions using odds ratios. *Statistics in Medicine* 2014; **33**:2275–2287.
33. Efthimiou O, Mavridis D, Riley RD, Cipriani A, Salanti G. Joint synthesis of multiple correlated outcomes in networks of interventions. *Biostatistics* 2015; **16**:84–97.
34. Debray TP, Schuit E, Efthimiou O, Reitsma JB, Ioannidis JP, Salanti G, Moons KG. Workpackage on behalf of G: an overview of methods for network meta-analysis using individual participant data: when do benefits arise? *Statistical Methods in Medical Research* 2016. 962280216660741.
35. Dias S, Welton NJ, Marinho VCC, Salanti G, Higgins JPT, Ades AE. Estimation and adjustment of bias in randomized evidence by using mixed treatment comparison meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2010; **173**:613–629.
36. Neuenschwander B, Branson M, Spiegelhalter DJ. A note on the power prior. *Statistics in Medicine* 2009; **28**:3562–3566.
37. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *The New England Journal of Medicine* 2000; **342**:1887–1892.
38. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine* 2000; **342**:1878–1886.
39. Viswanathan M, Berkman ND, Dryden DM, Hartling L. *Assessing Risk of Bias and Confounding in Observational Studies of Interventions or Exposures: Further Development of the RTI Item Bank*. Agency for Healthcare Research and Quality (US): Rockville (MD), 2013[*AHRQ Methods for Effective Health Care*].
40. Valentine JC, Thompson SG. Issues relating to confounding and meta-analysis when including non-randomized studies in systematic reviews on the effects of interventions. *Research Synthesis Methods* 2013; **4**:26–35.
41. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database of Systematic Reviews* 2014; **4** MR000034.
42. Welton NJ, Ades AE, Carlin JB, Altman DG, Sterne JC. Models for potentially biased evidence in meta-analysis using empirically based priors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2009; **172**:119–136.
43. Spiegelhalter D, Abrams KR, Myles JP: *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley & Sons, Ltd: Chichester, UK, 2003. DOI: 10.1002/0470092602.
44. Ibrahim JG, Chen M-H. Power prior distributions for regression models. *Statistical Science* 2000; **15**:46–60.

45. Ibrahim JG, Chen M-H, Gwon Y, Chen F. The power prior: theory and applications. *Statistics in Medicine* 2015; **34**:3724–3749.
46. Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* 2014; **70**:1023–1032.
47. Mutsvari T, Tytgat D, Walley R. Addressing potential prior-data conflict when using informative priors in proof-of-concept studies. *Pharmaceutical Statistics* 2016; **15**:28–36.
48. Prevost TC, Abrams KR, Jones DR. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Statistics in Medicine* 2000; **19**:3359–3376.
49. Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research* 2001; **10**:277–303.
50. Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials. *Medical Decision Making* 2013; **33**:641–656.
51. Salanti G, Del Giovane C, Chaimani A, Caldwell DM, Higgins JPT. Evaluating the quality of evidence from a network meta-analysis. *PLoS ONE* 2014; **9** e99682.
52. Krahn U, Binder H, König J. A graphical tool for locating inconsistency in network meta-analyses. *BMC Medical Research Methodology* 2013; **13**:35.
53. Furukawa TA, Miura T, Chaimani A, Leucht S, Cipriani A, Noma H, Mitsuyasu H, Kanba S, Salanti G. Using the contribution matrix to evaluate complex study limitations in a network meta-analysis: a case study of bipolar maintenance pharmacotherapy review. *BMC Research Notes* 2016; **9**:218.
54. Rücker G, Schwarzer G, Krahn U, König J: Netmeta: network meta-analysis using frequentist methods, 2015. *Available at:* http://www.r-Project.org.
55. Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G. Graphical tools for network meta-analysis in STATA. *PLoS ONE* 2013; **8** e76654.
56. Jackson D, White IR, Riley RD. Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Statistics in Medicine* 2012; **31**:3805–3820.
57. StataCorp. *Stata Statistical Software: Release 13*. StataCorp LP: College Station, TX, 2013.
58. White IR. Network meta-analysis. *Stata Journal* 2015; **15**:951–985.
59. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS &Ndash; a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**:325–337.
60. Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: evolution, critique and future directions. *Statistics in Medicine* 2009; **28**:3049–3067.
61. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology* 1997; **50**:683–691.
62. Almalla M, Schröder J, Pross V, Marx N, Hoffmann R. Paclitaxel-eluting balloon versus everolimus-eluting stent for treatment of drug-eluting stent restenosis. *Catheterization and Cardiovascular Interventions* 2014; **83**:881–887.
63. Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *Journal of Clinical Epidemiology* 2011; **64**:163–171.

Supporting information

Additional supporting information may be found online in the supporting information tab for this article.