

Combining individual patient data from randomized and non-randomized studies to predict real-world effectiveness of interventions

Statistical Methods in Medical Research

2022, Vol. 31(7) 1355–1373

© The Author(s) 2022



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/09622802221090759

journals.sagepub.com/home/smm

Michael Seo^{1,2} , Thomas PA Debray^{3,4} , Yann Ruffieux¹ ,
Sandro Gsteiger⁵ , Sylwia Bujkiewicz⁶, Axel Finckh⁷,
Matthias Egger^{1,8}, and Orestis Efthimiou^{1,9} 

Abstract

Meta-analysis of randomized controlled trials is generally considered the most reliable source of estimates of relative treatment effects. However, in the last few years, there has been interest in using non-randomized studies to complement evidence from randomized controlled trials. Several meta-analytical models have been proposed to this end. Such models mainly focussed on estimating the average relative effects of interventions. In real-life clinical practice, when deciding on how to treat a patient, it might be of great interest to have personalized predictions of absolute outcomes under several available treatment options. This paper describes a general framework for developing models that combine individual patient data from randomized controlled trials and non-randomized study when aiming to predict outcomes for a set of competing medical interventions applied in real-world clinical settings. We also discuss methods for measuring the models' performance to identify the optimal model to use in each setting. We focus on the case of continuous outcomes and illustrate our methods using a data set from rheumatoid arthritis, comprising patient-level data from three randomized controlled trials and two registries from Switzerland and Britain.

Keywords

Real-world effectiveness, individual patient data, non-randomized studies, network meta-analysis, efficacy-effectiveness gap

Introduction

Randomized clinical trials (RCTs) and meta-analyses (MAs) of clinical trials are usually thought to be the most reliable source of evidence to evaluate the effects of medical interventions.¹ However, RCTs are often carried out in specific settings and apply strict inclusion and exclusion criteria, which leads to results that may not be generalizable to routine

¹Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

²Graduate School for Health Sciences, University of Bern, Bern, Switzerland

³Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

⁴Smart Data Analysis and Statistics B.V., Utrecht, The Netherlands

⁵Pharmaceuticals Division, Global Access, F. Hoffmann-La Roche, Basel, Switzerland

⁶Biostatistics Research Group, Department of Health Sciences, University of Leicester, Leicester, UK

⁷Division of Rheumatology, University Hospitals of Geneva, Geneva, Switzerland

⁸Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

⁹Department of Psychiatry, University of Oxford, Oxford, UK

Corresponding author:

Michael Seo, Graduate School for Health Sciences, University of Bern, Bern, Switzerland.

Email: swj8874@gmail.com

practice.² This situation leads to the so-called efficacy-effectiveness gap,^{3,4} where ‘efficacy’ refers to the performance of a medical intervention under experimental conditions, while ‘effectiveness’ relates to what is achieved in everyday clinical practice.

In the last few years, there has been growing interest in utilizing ‘real-world’ data from non-randomized studies (NRSs), to complement evidence from RCTs in medical decision-making, aiming to bridge the efficacy-effectiveness gap.^{5–8} The increased interest in NRSs was accompanied by the development of various meta-analytical modelling approaches,^{9–12} and by extending methods to network meta-analysis (NMA) and individual patient data (IPD) MA.^{13–15} So far, methods have primarily focussed on estimating relative treatment effects. When making (shared) decisions¹⁶ on how to treat patients in real-world settings, the prediction of (possibly several) outcomes under all competing treatments for individual patients or specific groups of patients is of great interest. In this context, jointly synthesizing IPD from RCTs and non-randomized, real-world studies in a prediction framework is promising. This idea was previously explored by Didden et al.,¹⁷ who aimed to predict the average, population-level, real-world effectiveness of a treatment pre-launch (i.e. before a treatment becomes available for clinical practice). However, to the best of our knowledge, there are no methods that combine patient-level data from multiple studies, randomized and observational, to provide personalized predictions for various treatments for patients treated in real-world settings.

We set out to fill this gap and propose a general prediction framework. We extend prediction modelling methods to an NMA setting¹⁸ to synthesize evidence on multiple interventions, utilizing IPD from randomized and NRSs. We propose a range of modelling approaches for predicting real-world outcomes and describe methods for selecting the optimal model to use. We focus on two-stage models, where at the first stage each study is analysed separately, and at the second stage the study-specific results are meta-analysed. Although one-stage approaches to IPD-MA and IPD-NMA are generally considered more flexible,¹⁹ a two-stage approach is often necessary for practice, due to data restrictions, e.g. when data reside on different servers. We illustrate how to implement our framework, using a real example in rheumatoid arthritis (RA), combining data from multiple trials and registries.

A clinical example in RA

We used IPD from three RCTs and two NRS (based on disease registries) on patients diagnosed with RA, a chronic inflammatory disease characterized by progressive damage of joints.²⁰ Several drugs can be prescribed for the treatment of RA. In this example, we focus on three: conventional synthetic disease-modifying anti-rheumatic drugs (DMARDs), combination therapy with rituximab (RTX) + DMARDs, and a combination of tocilizumab (TCZ) + DMARDs. Note that RTX and TCZ are also DMARDs, but they are biologic in contrast to conventional DMARDs such as methotrexate, which we refer to as DMARDs. The outcome of interest was the Disease Activity Score 28 (DAS28),²¹ which is continuous and ranges from 0 to 10. A lower score indicates lower disease activity.

One RCT (REFLEX) compared RTX + DMARDs versus DMARDs and included 517 patients. The other two RCTs (TOWARD and TOWARD2) compared TCZ + DMARDs versus DMARDs, and included 1216 and 791 patients, respectively. In all three RCTs, DAS28 was at baseline and after 6 months. For the two NRS, we used the measurement of DAS28 at 6 month (or the closest available measurement, within ± 3 months) after the start of the new drug. The British Society for Rheumatology Biologics Register in Rheumatoid Arthritis (BSRBR-RA) and the Swiss Clinical Quality Management in Rheumatoid Arthritis (SCQM) registries provided observational data.^{22,23} These included 2057 and 1069 patients, respectively. Established in 2001, BSRBR-RA is one of the largest studies looking at the long-term safety of new drugs prescribed for RA. It was launched when anti-tumor necrosis factor (TNF) therapies became available.²¹ Similarly, SCQM is a Swiss registry established in 1997 and allowed rheumatologists to follow their RA patients to improve outcomes.²² In addition to baseline DAS28, nine patient-level covariates were available from both RCTs and NRS at baseline: gender, age, disease duration, body mass index, baseline rheumatoid factor, number of previous DMARDs and anti-TNF agents, baseline health assessment questionnaire disability index, and baseline erythrocyte sedimentation rate (ESR). We provide additional details on the studies in Tables 1–3 of the Appendix. Figure 1 shows the network graph and overview of data sets.

Methods

Overview

Below, we propose a statistical framework to facilitate personalized medicine when IPD from NRS and RCTs are available. This framework can be used to build a multivariable model that predicts (absolute) patient-level, real-world outcomes for a range of different treatments for a given disease. We assume that the available data are IPD from a set of RCTs and NRS that compare two or more treatments for the same disease. The final product of our analyses will be a model where the input

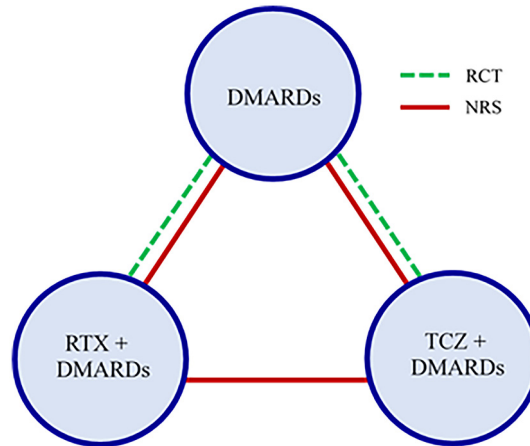


Figure 1. Network graph for the rheumatoid arthritis example. Green lines show RCTs and red lines show NRS. There were three two-armed RCTs; two of them compared TCZ + DMARDs vs DMARDs and one compared RTX + DMARDs vs DMARDs. There were two NRS, which included all three drugs. Abbreviations: DMARDs: Conventional disease-modifying anti-rheumatic drugs; RTX: rituximab; TCZ: tocilizumab; RCT: randomized controlled trial; NRS: non-randomized study.

is a vector of patient-level covariates (\mathbf{x}), and the output is the predicted outcome under each treatment (t). In what follows, we focus on the case of a continuous outcome and we limit our investigation to models with linear functions of the covariates. The model is as follows:

$$y_{\text{pred}}(\mathbf{x}, t) = \alpha + \boldsymbol{\beta}^T \mathbf{x} + \boldsymbol{\gamma}_{tA}^T \mathbf{x} + \delta_{tA} \quad (1)$$

where A is a reference treatment and \mathbf{x} is a vector of patient-level covariates that may impact the outcome, possibly interacting with treatment. Without loss of generality, we assume all continuous covariates to be standardized, to facilitate model fitting. Next, α is the predicted outcome under treatment A when $\mathbf{x} = 0$; $\boldsymbol{\beta}$ encompasses the prognostic ability of the covariates; $\boldsymbol{\gamma}_{tA}$ expresses the effect modification (i.e. treatment-covariate interactions) of treatment t versus A ; and δ_{tA} is the relative treatment effect t versus A for $\mathbf{x} = 0$. We set $\boldsymbol{\gamma}_{AA}$ and δ_{AA} equal to zero. If a non-linear relationship is suspected, we can expand Equation (1) to include fractional polynomials or cubic-spline terms. Such extensions are straightforward within the framework described below but are beyond the scope of this paper.

We describe a range of meta-analytical two-stage approaches that can be used to estimate the parameters in the prediction model of Equation (1). First, we analyse each study separately. Second, we meta-analyse the study-specific estimates, to estimate all parameters needed for Equation (1).

We focus on Bayesian approaches because of their flexibility and ease of propagating uncertainty. We describe three generic approaches for building prediction models that combine IPD from studies of variable design.

- *Approach I:* We only use the IPD from a single NRS, i.e. without using MA.
- *Approach II:* We combine the IPD from all studies (NRS and RCTs) using a two-stage NMA, considering three different implementations. In approach IIa, we do not account for differences in study design. In approach IIb we employ shrinkage methods. In approach IIc, we further calibrate²⁴ the intercept and main effects of covariates to target a specific patient group, using data from a single NRS that reflects this population.
- *Approach III:* We combine the IPD from NRS and RCTs using a two-stage NMA and adopt a weighting scheme to account for different study designs (IIIa). Again, we allow for re-calibrating the intercept term and main effects of covariates to target a specific patient group (IIIb).

As an additional, simplified method, we combine all studies as in IIa but we do not include any treatment-covariate interactions, i.e. we set $\boldsymbol{\gamma} = 0$. We call this Approach IV. After describing the various approaches in detail, we discuss ways of assessing their performance. Table 1 summarizes each approach, briefly describing the model and citing equations used.

Approach I: Analysis of a single study, reflecting the target population

Let us assume that our aim is to predict outcomes for a real-world population from which we have a representative sample, including patient-level data (e.g. from a specific registry). We can build a prediction model using these data alone (e.g. the model of

Equation (1)) and disregard all other sources of information. This approach is relatively common in the literature, although (a) usual applications do not account for multiple interventions, and (b) these models are often based on IPD from a single RCT.

To enhance the predictive performance of the model, we can incorporate shrinkage in the model estimation. Shrinkage methods are known to improve prediction accuracy,^{24,25} while it has been recently shown that they can be useful in a MA setting, when aiming to estimate patient-level treatment effects.²⁶ We are particularly interested in shrinking treatment-covariate interactions because information on effect modification is important, as it may substantially affect decision making.

There are many different frequentist as well as Bayesian shrinkage methods we could use. For example, for Bayesian shrinkage, we can penalize the coefficients of the effect modification (treatment-covariate interactions, i.e. γ parameters in Equation (1)) using a Laplace prior distribution. When the scale parameter of the Laplace prior distribution is assigned its own prior distribution, the model is often referred to as a Bayesian LASSO.^{27,28} The prior for effect modifiers of treatment W versus the reference A is given by:

$$\pi(\gamma_{WA}|\lambda) = \prod_{k=1}^{n_{cov}} \frac{\lambda}{2} e^{-\lambda|\gamma_{k,WA}|} \tag{2}$$

where n_{cov} denotes the total number of covariates, $\gamma_{k,WA}$ is the regressor for effect modification of the k th covariate, and λ is the shrinkage parameter that controls how much estimates are shrunk to zero. There are several ways to select the shrinkage parameter, e.g. we can use k -fold cross validation or choose the value via Bayes factor.²⁹ We hereby chose to treat it as random and assign a non-informative hyperprior. Details on the hyperprior are given in the ‘Fitting the models’ section. Of note, Bayesian LASSO does not strictly perform variable selection, since covariates are not shrunk all the way to zero. As discussed above, we used penalization only for γ . Alternatively, we could also penalize prognostic factors, i.e. the β parameters as well. For further discussion on possible extensions of this model we refer the reader to the Discussion section.

Approach II: Design-naïve NMA

In this approach, we use all studies to fit a two-stage IPD NMA model.³⁰ We disregard the information on study design, i.e. we treat RCTs and NRS the same. Thus, we use all studies to inform parameters $\beta, \gamma_{tA}, \delta_{tA}$ in Equation (1). The α parameter of this model, which denotes the predicted outcome for zero value of the covariates under the reference treatment A , will only be informed by studies that included this treatment. This means that the choice of reference treatment is potentially important (while the choice of reference is arbitrary in the standard NMA model.¹⁸) An obvious choice A is the treatment that is given to most patients across the available data sets.

Let us assume that patient i was included in the study j , received treatment t_{ij} , and that for this patient the observed outcome of interest was y_{ij} , measured on a continuous scale. Also, we assume that for this patient we have \mathbf{x}_{ij} covariates. Let us also assume that the study j had only two treatment arms, W and A (the reference). Under these definitions, the first stage model is as follows:

$$\begin{aligned} y_{ij} &\sim N(m_{ij}, \sigma_j^2) \\ m_{ij} &= a_j + \mathbf{b}_j^T \mathbf{x}_{ij} + \mathbf{c}_{j,t_{ij}A}^T \mathbf{x}_{ij} + d_{j,t_{ij}A} \\ \sigma_j^2, a_j, \mathbf{b}_j, \mathbf{c}_{j,WA}, d_{j,WA} &\sim (\text{vague prior distributions}) \end{aligned} \tag{3}$$

where we set $\mathbf{c}_{j,AA} = d_{j,AA} = 0$; m_{ij} denotes the expected outcome (linear predictor); σ_j^2 is the study-specific variance of the outcome; a_j is the study-specific intercept; \mathbf{b}_j is the vector of study-specific regression coefficients for the main effects of the covariates; $\mathbf{c}_{j,WA}$ is the vector of coefficients for effect modification (i.e. treatment-covariate interaction) for study j and treatment W relative to A ; and $d_{j,WA}$ is the treatment effect at $\mathbf{x} = 0$, for study j and treatment W relative to A . Prior distributions for these parameters are described in more detail in the ‘Fitting the models’ section. Thus, at the first stage of the model we estimate $\hat{\theta}_j = (\hat{a}_j, \hat{\mathbf{b}}_j, \hat{\mathbf{c}}_{j,WA}, \hat{d}_{j,WA})$ and the corresponding variance-covariance matrix $\hat{\mathbf{S}}_j$.

These estimates contribute to the likelihood of the second stage of the MA, via a multivariate random-effects model:

$$\begin{aligned} \hat{\theta}_j &\sim N(\xi_j, \hat{\mathbf{S}}_j) \\ \xi_j &\sim N((\alpha, \beta, \gamma_{WA}, \delta_{WA}), \mathbf{\Sigma}) \\ (\alpha, \beta, \gamma_{WA}, \delta_{WA}, \mathbf{\Sigma}) &\sim (\text{vague prior distributions}) \end{aligned} \tag{4}$$

where $\mathbf{\Sigma}$ is a between-study (i.e. random effects) variance-covariance matrix. Note that when using the IPD NMA model, we are typically interested in estimating relative effects (i.e. γ, δ), while α is usually set to be study-specific and is

Table 1. Overview of the different modelling approaches presented in this paper.

Approach	Description	Prediction for a new patient with covariates \mathbf{x} , treatment \mathbf{t}	Meta-analysis	α , β in prediction model estimated from:	γ , δ in prediction model estimated from:	Shrinkage at the first stage	Formulas used at the first stage	Formulas used at the second stage
I	We only use a single NRS.							
IIa	First stage: fit a model in each study separately. Second stage: fit a design-naive NMA.		✗ ✓	A single NRS All RCTs and NRS	A single NRS All RCTs and NRS	✓ ✗	(2) and (3) (3)	- (5) or (6)
IIb	Same as IIa, but at first stage we use penalized estimation.		✓	All RCTs and NRS	All RCTs and NRS	✓	(2) and (3)	(5) or (6)
IIc	Same as IIb, but intercept and main effects (α , β) are estimated from a single NRS.		✓	A single NRS	All RCTs and NRS	✓	(2) and (3)	(5) or (6)
IIIa	First stage: as per IIb Second stage: use weights according to study design.		✓	All NRS	A combination of RCTs and NRS	✓	(2) and (3)	(7) and (8)
IIIb	Same as IIIa, but intercept and main effects are estimated from a single NRS.		✓	A single NRS	A combination of RCTs and NRS	✓	(2) and (3)	(7)
IV	Same as IIa, but without effect modifiers	$y_{\text{pred}}(\mathbf{x}, \mathbf{t}) = \sigma + \beta^T \mathbf{x} + \delta_{tA}$	✓	All RCTs and NRS	All RCTs and NRS	✗	(9)	(10)

Abbreviations: RCT: randomized clinical trial. NRS: non-randomized study. NMA: network meta-analysis.

considered a nuisance parameter. Here, however, we want to predict absolute outcomes; thus, we also require estimates of α, β . The estimation of Σ might be difficult, especially for a few studies or many covariates. In this case, we may simplify the model by assuming α, β , and γ_{WA} to be common (fixed), i.e. independent of the study.

Furthermore, a typical assumption in NMA is that the between-study variance (τ^2) of the random effects is the same across treatment contrasts.¹⁸ Then, the second stage model can be written as follows:

$$\begin{aligned}\hat{\theta}_j &\sim N((\alpha, \beta, \gamma_{WA}, \theta_{j,WA}^{(\delta)}), \hat{S}_j) \\ \theta_{j,WA}^{(\delta)} &\sim N(\delta_{WA}, \tau^2) \\ (\alpha, \beta, \gamma_{WA}, \delta_{WA}, \tau^2) &\sim (\text{vague prior distributions})\end{aligned}\quad (5)$$

For multi-arm studies and for studies that do not include reference treatment A , the models need adaptation; we provide details in Section 2 of the Appendix.

Again, it may be difficult to estimate heterogeneity (τ^2) of the treatment effect δ when only a few studies are available. This is the case for the RA example described in the ‘A clinical example in RA’ section. One solution would be to use external information to create an informative prior distribution for τ^2 . For example, for the case of binary outcomes, Turner et al.³¹ proposed empirical distributions that can be used in Bayesian MAs. A further simplification would be to assume common δ :

$$\begin{aligned}\hat{\theta}_j &\sim N((\alpha, \beta, \gamma_{WA}, \delta_{WA}), \hat{S}_j) \\ (\alpha, \beta, \gamma_{WA}, \delta_{WA}) &\sim (\text{vague prior distributions})\end{aligned}\quad (6)$$

We call Approach IIa the combination of Equation (3) and Equation (5) or (6) to estimate the parameters of Equation (1). Approach IIb is an extension of IIa, where we use the shrinkage method (i.e. Bayesian LASSO) at the first stage.

Lastly, we can further extend Approach IIb by calibrating the intercept and main effects of covariates to target a specific real-world population. A motivation for this is that data sampled from the patient population of interest might be the best source of evidence for predicting the reference treatment outcome. Conversely, estimates of the intercept term obtained from RCTs might be less representative for the target population because RCT patients are selected and because RCTs are performed in highly controlled settings.

In detail, we use Approach I to estimate α, β using only data obtained from a study that reflects the target population, and Approach IIb, using all data to estimate γ, δ . We then use Equation (1) to make predictions for individual patients. We call this Approach IIc. To summarize, in Approach IIc we use study-specific estimates of α, β and pooled relative effects, i.e. γ, δ . Note that we fit the full model within each study separately, and then pool all parameters (including all main effects and interactions) in the second stage of the model across studies. Table 1 summarizes all the different flavours of Approach II.

Approach III: Design-adjusted analysis

For Approach III, when aggregating parameters at the second stage, we use a weighting scheme for studies of a different design.¹⁵ Specifically, we weight the first-stage estimates of relative treatment effects from Approach IIb according to the corresponding study’s design.

In this approach, the variance of the estimates of treatment effects and effect modification obtained from NRS j (i.e. the variance of $\hat{c}_{j,WA}, \hat{d}_{j,WA}$) is inflated after dividing by a factor w_j , with $0 < w_j < 1$. By doing so, we effectively decrease the impact of NRS in the estimation of all relative treatment effects. Setting $w_j=0$ corresponds to completely disregarding estimates of relative treatment effects obtained from NRS.

This approach’s motivation is that RCTs are usually thought to be the most reliable sources of information for relative treatment effects because randomization helps us avoid issues related to confounding. Moreover, since we aim to predict outcomes for real-world populations, estimates for the model’s intercept (α) and the main effects of the covariates (β) are aggregated only using NRS. The second stage model is as follows:

$$\begin{aligned}(\hat{c}_{j,WA}, \hat{d}_{j,WA}) &\sim \begin{cases} N\left(\gamma_{WA}, \delta_{WA}, \frac{\hat{S}_j}{w_j}\right), & \text{if study } j \text{ is an NRS} \\ N((\gamma_{WA}, \delta_{WA}), \hat{S}_j), & \text{if study } j \text{ is an RCT} \end{cases} \\ (\gamma_{WA}, \delta_{WA}) &\sim (\text{vague prior distributions})\end{aligned}\quad (7)$$

and

$$\begin{aligned}(\hat{a}_j, \hat{b}_j) &\sim N((\alpha, \beta), \hat{S}_j), \quad \text{if study } j \text{ is a NRS} \\ (\alpha, \beta) &\sim (\text{vague prior distributions})\end{aligned}\tag{8}$$

For choosing the weights we can consider e.g. study quality, or how similar is the healthcare system of an NRS from a specific country as compared to that of a target population;¹⁵ in practice, a range of values can be used, and model performance measures can be assessed (see next section) to decide on the optimal weights. See also the Discussion section for some additional considerations regarding weights. We call this Approach IIIa. Finally, instead of aggregating NRS to estimate study intercept and main effects, we can calibrate to target a population as in Approach IIc. We call this approach IIIb.

Approach IV: Design-naïve NMA without effect modifiers

For Approach IV, we fit the similar design-naïve NMA, but without including effect modifiers (i.e. setting $\gamma = 0$). This may serve as a sensitivity analysis, and by comparing its results with the previous methods we might obtain insights on the extent of heterogeneity of treatment effects. The linear predictor is now

$$m_{ij} = a_j + \mathbf{b}_j^T \mathbf{x}_{ij} + d_{j,t_{ij}A}\tag{9}$$

and the multivariate MA is

$$\begin{aligned}\hat{\theta}_j &\sim N(\boldsymbol{\psi}_j, \hat{S}_j) \\ \boldsymbol{\psi}_j &\sim N((\alpha, \beta, \delta_{WA}), \boldsymbol{\Sigma}) \\ (\alpha, \beta, \delta_{WA}, \boldsymbol{\Sigma}) &\sim (\text{vague prior distributions})\end{aligned}\tag{10}$$

$$\text{where } \hat{\theta}_j = (a_j, b_j, d_{j,t_{ij}A})$$

Assessing the performance of the prediction models

A common practice to measure a prediction model's performance for a continuous outcome is through mean squared error (MSE) and bias:

$$\text{MSE} = \frac{1}{N} \sum_{ij} (\hat{y}_{ij} - y_{ij})^2, \quad \text{Bias} = \frac{1}{N} \sum_{ij} (\hat{y}_{ij} - y_{ij})\tag{11}$$

where N is the total number of patients in a given study of interest, \hat{y}_{ij} is the predicted outcome using equation (1) and y_{ij} is the observed outcome. Values close to zero indicate good performance. Note that since the focus is on predicting real-world outcomes, metrics should be evaluated on NRS data.

We can also calculate the coefficient of determination (R-squared) to evaluate a model's performance for a given study of interest. We do that by using only y_{ij} and \hat{y}_{ij} from each model:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}\tag{12}$$

where $SS_{\text{tot}} = \sum_{i,j} (y_{ij} - \bar{y})^2$ is the total sum of squares, \bar{y} is the average observed outcome, and $SS_{\text{res}} = \sum_{i,j} (\hat{y}_{ij} - y_{ij})^2$ is the residual sum of squares. We can also fit a regression line to inspect the agreement between observed and predicted patient outcomes, i.e.

$$y_{ij} = \eta_0 + \sum_{k=1}^K \eta_k \hat{y}_{ij} I(t_{ij} = k)\tag{13}$$

where K is the total number of treatments and $I(t_{ij} = k)$ is an indicator function that equals 1 if the treatment assigned to a patient i in study j was k , and 0 otherwise. This 'calibration line' compares the observed (y_{ij}) with the predicted outcome (\hat{y}_{ij}) across different treatment groups. Having η_0 close to 0 and η_1, \dots, η_K values close to 1 indicate good performance of the prediction models.

The performance measures in equations (11), (12), and (13) inform us of our models' overall predictive ability. However, when deciding on how to treat, we are mainly interested in selecting the best treatment for each patient.

Thus, a prediction model that would be useful for deciding between competing treatments should also be well calibrated in terms of estimated treatment benefit. To achieve this, we reshape the calibration line using the predicted benefit (i.e. the difference between predicted outcomes under different treatments). For networks with treatments 1, 2, 3, ... for each of the competing models we fit the line

$$y_{ij} = \kappa_0 + \kappa_1 \hat{y}_{ij,1} + \kappa_2 (\hat{y}_{ij,2} - \hat{y}_{ij,1}) I(t_{ij} = 2) + \kappa_3 (\hat{y}_{ij,3} - \hat{y}_{ij,1}) I(t_{ij} = 3) + \dots \quad (14)$$

where $\hat{y}_{ij,1}, \hat{y}_{ij,2}, \hat{y}_{ij,3}, \dots$ are the predicted outcomes of a patient under treatments 1, 2, 3, ... respectively, under this model. Similarly, having κ_0 close to 0 and $\kappa_1, \kappa_2, \kappa_3, \dots$ values close to 1 indicate good performance of the prediction models.

Cross validation

For each competing model, we can calculate the performance metrics discussed above using internal as well as an internal-external cross validation.^{32,33}

For internal validation, we use all available data to develop the models. Then, given that we aim to predict real-world outcomes, we use the NRS to test all models, i.e. to compare predictions with observations using the measures described in the previous section. To correct for optimism,²⁴ we can calculate optimism-corrected performance using bootstrapping as discussed by Steyerberg.²⁴ More details on this procedure are given in Section 2.2 of the Appendix. An internal validation procedure will inform us on which model performs best for each specific setting, provided we stratify the procedure by study.

For internal-external cross-validation, we exclude one NRS from the analysis and use the rest of the data to train the models. We then use the left-out NRS to make predictions and compare them with observations. Finally, we cycle through all available NRS. To follow this approach, we need data from multiple NRS to be available. If this is not the case, we may split the single NRS data in a meaningful non-random way, e.g. by clinic, region, or any other clustering variable in our data set. This internal-external validation procedure can potentially provide insights on which model might perform better in new settings.

The internal-external cross-validation via leave-one-study-out cannot be combined with Approach I, as this analysis only uses a single study to train the model. In this case, we can instead use all NRS except the left-out one to develop the model. For instance, for the RA case study, we only used the British registry to train the model and the Swiss registry to make predictions, and vice versa. Moreover, for Approach IIc and IIIb, when using internal-external cross-validation, we use all NRS except the left-out to estimate intercept and main effects of covariates, i.e. we do not use RCTs.

Implementation details

Below we provide some details on implementing the described models in the RA example.

Standardization of covariates

To use penalized estimation in the analysis of study j we need to standardize variables. This is necessary to ensure that penalization is equally applied to all regressors.²⁵ Standardizing means transforming each covariate in each study by subtracting the study-specific mean from the covariate and dividing the result with the covariate's study-specific standard deviation. However, standardizing makes it difficult to MA results from multiple studies coherently, given that in each study the covariates are transformed differently. Thus, before aggregating the first stage results at the second stage, we need to revert coefficients to their natural scale. The mathematical details of how to do this are provided in Section 2.3 of the Appendix.

Imputation of missing data

Gelman et al.³⁴ recommended the following approach to handle missing data in Bayesian analyses via Markov Chain Monte Carlo (MCMC): (a) create m multiply imputed data sets; (b) analyse each imputed data set separately; (c) combine the m posterior draws, i.e. by mixing the corresponding draws.

We use R package `mice`³⁵ to impute the missing covariates to create $m = 20$ multiply imputed data sets.³⁶ When imputing we used information from covariates, treatment, covariate-treatment interactions, and outcomes, but we did not impute the missing outcomes for model development. Imputation was done in each study separately using the method of predictive mean matching.

Fitting the models

All analyses were carried out in R³⁷ using `rjags`.³⁸ For all models, DMARDs were the reference treatment. When fitting the first stage models, we used 20 imputed data sets and ran 3 chains of 10,000 iterations each, with 1000 burn-in. For the second stage models, we ran 3 chains of 200,000 iterations with 20,000 burn-in. We assessed convergence using the Gelman and Rubin diagnostics.³⁹

For all models, we used a vague prior distribution for the standard deviation of continuous outcomes ($\sigma_j \sim \text{Uniform}(0, 5)$). For regression parameters of stage one (i.e. α , β , γ_{WA} , and δ_{WA}), we used a Normal(0, $\sigma^2 = 1000$) distribution. When applying Bayesian LASSO in stage one, a vague prior distribution was placed on the scale parameter for Laplace prior ($\lambda^{-1} \sim \text{Uniform}(0, 5)$). We tested the sensitivity of results to the selection of prior distribution on λ . For Approach III, we used weights 0.25 and 0.5.

For calculating optimism-corrected performance, 200 bootstrap samples were drawn. The R codes used for fitting all models are available at <https://github.com/MikeJSeo/phd/tree/master/ra>.

Results

The parameter estimates from the first-stage analysis for each study are shown in Tables 4, 5, 6, and 7 in the Appendix. There were some important differences between RCTs and NRS, especially in the case of the Swiss registry. The estimated intercept term for this study was much smaller than for the RCTs, probably owing to the fact that the Swiss registry had the lowest average DAS28 score; see Table 3 in the Appendix. Moreover, the average relative treatment effects of the biologic treatments versus DMARDs estimated in the registries were smaller than in the RCTs. These differences might raise concerns about the direct applicability of the RCT findings in patients found in real-world settings. We might hypothesize that these differences were due to residual confounding, model misspecification (e.g. omission of non-linear or interaction terms), or due to the more general ‘efficacy-effectiveness gap’ described in the Introduction.

Furthermore, there was only very weak evidence of heterogeneous treatment effects (i.e. effect modification) across all studies. Results from the second-stage analysis are shown in Table 8 of the Appendix. As expected, given the first stage results, methods that gave more weight to the RCTs in the estimation of relative effects provided larger effects. Next, we assessed all models’ performance using the measures described in the ‘Assessing the performance of the prediction models’ section and following the two cross-validation approaches described in the ‘Cross validation’ section.

We first discuss the results from the internal validation. In Table 2, we give results in terms of MSE and Bias, and in Table 3 results for the calibration lines and R-squared value. Figures 1 and 2 in the Appendix show the calibration plots. Overall, we saw that for most approaches the bias was rather small in clinical terms. MSE was around 1.5 for all models; to bring it to the same scale as the outcome we can calculate the root of MSE, i.e. 1.22, deemed to be small-to-moderate in clinical terms. R-squared was moderate (around 0.30 for all approaches). Moreover, we see that most models had similar performance. For the Swiss registry, Approach I performed slightly better for most measures of performance. For the British registry, approach IIIa was overall the best, especially in measures of calibration. However, differences with other methods, including Approach I, were not very pronounced. This showed that, when predicting real-world outcomes, utilizing data from multiple studies and using advanced analysis methods brought no benefit to the Swiss and small benefits to the British registry.

Next, we discuss results from the internal–external validation. This procedure gave us some insight into the models’ expected performance when applied in new settings. Results are summarized in Tables 4, 5, and Figures 2, 3. Overall, calibration lines’ intercepts were far from zero for most models, and R-squared values were quite low, indicating low performance. Removing Swiss registry data from the training set and using it for validation resulted in models IIa, IIb, and IIIb performing best. Likewise, when we left the British registry out, we saw that model IIb and IV performed best. Approaches IIb and IV gave almost the same results and were overall the best for both registries. They had the lowest overall MSE, relatively higher R-squared value and lower bias for all treatment arms and were slightly better calibrated for both the absolute outcome and treatment benefit compared to the other models. This showed that utilizing information from RCTs can help us better predict real-world outcomes compared to using only data from NRS. It also showed that in this data set there was little evidence of an effect modification. Thus, if we aim to make predictions about patients in a new setting (e.g. another country), for which no data are currently available, we would recommend Approaches IIb or IV. See also the Discussion section on additional considerations regarding generalizability.

Figure 4 summarizes the MSE and bias results for both internal and internal-external validation. In Table 9 of the Appendix we show results after correcting for optimism. We saw that Approach I had slightly larger optimism, as compared to other methods. However, these optimism-corrected results did not materially change the conclusion drawn from the main analyses. This is because optimism was generally quite small for all models. This was expected, since

Table 2. Internal validation results of bias and MSE for different approaches.

Data set	Treatment arms	Performance metric	Approach I	Approach IIa	Approach IIb	Approach IIc	Approach IIIa w=0.25	Approach IIIa w=0.5	Approach IIIb w=0.25	Approach IIIb w=0.5	Approach IV	
SCQM	All arms	MSE	1.44	1.54	1.59	1.46	1.54	1.53	1.47	1.47	1.62	
		Bias	0.10	0.22	0.25	0.05	0.18	0.18	0.04	0.05	0.27	
	DMARDs	MSE	1.55	1.66	1.72	1.55	1.65	1.64	1.55	1.55	1.75	
		Bias	0.11	0.27	0.32	0.11	0.25	0.24	0.11	0.11	0.35	
	RTX+	MSE	0.94	1.05	1.03	0.98	1.04	1.04	0.99	0.98	1.04	
		Bias	0.11	0.12	0.10	-0.09	0.05	0.10	-0.13	-0.06	0.06	
	TCZ+	MSE	1.13	0.99	1.10	1.44	1.24	1.21	1.54	1.51	1.13	
		Bias	-0.16	-0.24	-0.36	-0.58	-0.47	-0.44	-0.66	-0.63	-0.35	
	BSRBR-RA	All arms	MSE	1.39	1.43	1.43	1.45	1.42	1.41	1.44	1.44	1.45
		Bias	0.13	0.11	0.07	0.19	0.09	0.05	0.18	0.20	0.05	
DMARDs	MSE	1.43	1.40	1.38	1.43	1.30	1.42	1.43	1.43	1.38		
	Bias	0.16	0.08	0.06	0.16	-0.03	0.08	0.16	0.16	0.02		
RTX+	MSE	1.28	1.28	1.30	1.28	1.82	1.30	1.27	1.29	1.34		
	Bias	0.11	0.09	0.04	0.14	0.13	0.01	0.14	0.19	0.00		
TCZ+	MSE	1.66	1.97	1.98	2.10	1.98	1.80	2.00	2.00	2.02		
	Bias	0.16	0.30	0.25	0.45	-0.30	0.13	0.38	0.40	0.31		

Abbreviations: MSE: mean squared error; DMARDs: disease-modifying anti-rheumatic drugs; RTX: rituximab; TCZ: tocilizumab.

Table 3. Internal validation results of the calibration lines and R-squared for different approaches.

Data set	Performance metric	Approach I	Approach IIa	Approach IIb	Approach IIc	Approach IIIa w=0.25	Approach IIIa w=0.5	Approach IIIb w=0.25	Approach IIIb w=0.5	Approach IV
SCQM	Calibration slope for outcome	$\eta_0 = 0.00$	$\eta_0 = 0.43$	$\eta_0 = 0.46$	$\eta_0 = 0.06$	$\eta_0 = 0.31$	$\eta_0 = 0.26$	$\eta_0 = 0.09$	$\eta_0 = 0.08$	$\eta_0 = 0.48$
		$\eta_1 = 0.97$	$\eta_1 = 0.81$	$\eta_1 = 0.79$	$\eta_1 = 0.95$	$\eta_1 = 0.85$	$\eta_1 = 0.86$	$\eta_1 = 0.94$	$\eta_1 = 0.95$	$\eta_1 = 0.78$
		$\eta_2 = 0.97$	$\eta_2 = 0.85$	$\eta_2 = 0.85$	$\eta_2 = 1.01$	$\eta_2 = 0.90$	$\eta_2 = 0.90$	$\eta_2 = 1.01$	$\eta_2 = 1.00$	$\eta_2 = 0.85$
		$\eta_3 = 1.06$	$\eta_3 = 0.92$	$\eta_3 = 0.95$	$\eta_3 = 1.24$	$\eta_3 = 1.05$	$\eta_3 = 1.07$	$\eta_3 = 1.26$	$\eta_3 = 1.25$	$\eta_3 = 0.93$
		$\kappa_0 = -0.01$	$\kappa_0 = 0.43$	$\kappa_0 = 0.43$	$\kappa_0 = 0.00$	$\kappa_0 = 0.26$	$\kappa_0 = 0.22$	$\kappa_0 = 0.01$	$\kappa_0 = 0.01$	$\kappa_0 = 0.43$
	Calibration slope for benefit	$\kappa_1 = 0.98$	$\kappa_1 = 0.81$	$\kappa_1 = 0.80$	$\kappa_1 = 0.97$	$\kappa_1 = 0.86$	$\kappa_1 = 0.87$	$\kappa_1 = 0.96$	$\kappa_1 = 0.96$	$\kappa_1 = 0.79$
		$\kappa_2 = 1.10$	$\kappa_2 = 0.57$	$\kappa_2 = 0.50$	$\kappa_2 = 0.68$	$\kappa_2 = 0.59$	$\kappa_2 = 0.66$	$\kappa_2 = 0.64$	$\kappa_2 = 0.71$	$\kappa_2 = 0.40$
		$\kappa_3 = 0.80$	$\kappa_3 = 0.66$	$\kappa_3 = 0.58$	$\kappa_3 = 0.59$	$\kappa_3 = 0.58$	$\kappa_3 = 0.59$	$\kappa_3 = 0.56$	$\kappa_3 = 0.57$	$\kappa_3 = 0.56$
		0.29	0.24	0.22	0.28	0.24	0.25	0.28	0.28	0.20
BSRBR-RA	Calibration slope for outcome	$\eta_0 = -0.46$	$\eta_0 = 0.29$	$\eta_0 = 0.27$	$\eta_0 = -0.07$	$\eta_0 = 0.05$	$\eta_0 = -0.03$	$\eta_0 = -0.03$	$\eta_0 = -0.08$	$\eta_0 = -0.36$
		$\eta_1 = 1.07$	$\eta_1 = 0.92$	$\eta_1 = 0.93$	$\eta_1 = 0.99$	$\eta_1 = 0.97$	$\eta_1 = 0.99$	$\eta_1 = 0.98$	$\eta_1 = 0.99$	$\eta_1 = 0.92$
		$\eta_2 = 1.08$	$\eta_2 = 0.91$	$\eta_2 = 0.93$	$\eta_2 = 0.98$	$\eta_2 = 0.99$	$\eta_2 = 1.01$	$\eta_2 = 0.98$	$\eta_2 = 0.97$	$\eta_2 = 0.91$
		$\eta_3 = 1.09$	$\eta_3 = 0.78$	$\eta_3 = 0.79$	$\eta_3 = 0.85$	$\eta_3 = 0.91$	$\eta_3 = 0.94$	$\eta_3 = 0.86$	$\eta_3 = 0.87$	$\eta_3 = 0.75$
		$\kappa_0 = -0.43$	$\kappa_0 = 0.26$	$\kappa_0 = 0.24$	$\kappa_0 = -0.14$	$\kappa_0 = -0.02$	$\kappa_0 = -0.08$	$\kappa_0 = -0.09$	$\kappa_0 = -0.10$	$\kappa_0 = -0.33$
	Calibration slope for benefit	$\kappa_1 = 1.07$	$\kappa_1 = 0.93$	$\kappa_1 = 0.94$	$\kappa_1 = 1.00$	$\kappa_1 = 0.99$	$\kappa_1 = 1.01$	$\kappa_1 = 0.99$	$\kappa_1 = 0.99$	$\kappa_1 = 0.92$
		$\kappa_2 = 1.04$	$\kappa_2 = 0.98$	$\kappa_2 = 1.01$	$\kappa_2 = 0.99$	$\kappa_2 = 0.96$	$\kappa_2 = 1.06$	$\kappa_2 = 0.98$	$\kappa_2 = 1.07$	$\kappa_2 = 0.90$
		$\kappa_3 = 1.05$	$\kappa_3 = 1.14$	$\kappa_3 = 1.13$	$\kappa_3 = 1.18$	$\kappa_3 = 1.06$	$\kappa_3 = 1.07$	$\kappa_3 = 1.14$	$\kappa_3 = 1.15$	$\kappa_3 = 1.16$
		0.39	0.37	0.37	0.36	0.38	0.38	0.37	0.37	0.36

These follow from linear regressions of the observed vs. predicted, as noted in equations (13) and (14). Subscript 0 refers to the intercept, 1 refers to DMARDs, 2 refers to RTX + DMARDs, and 3 refers to TCZ + DMARDs. Abbreviations: DMARDs: disease-modifying anti-rheumatic drugs; RTX: rituximab; TCZ: tocilizumab.

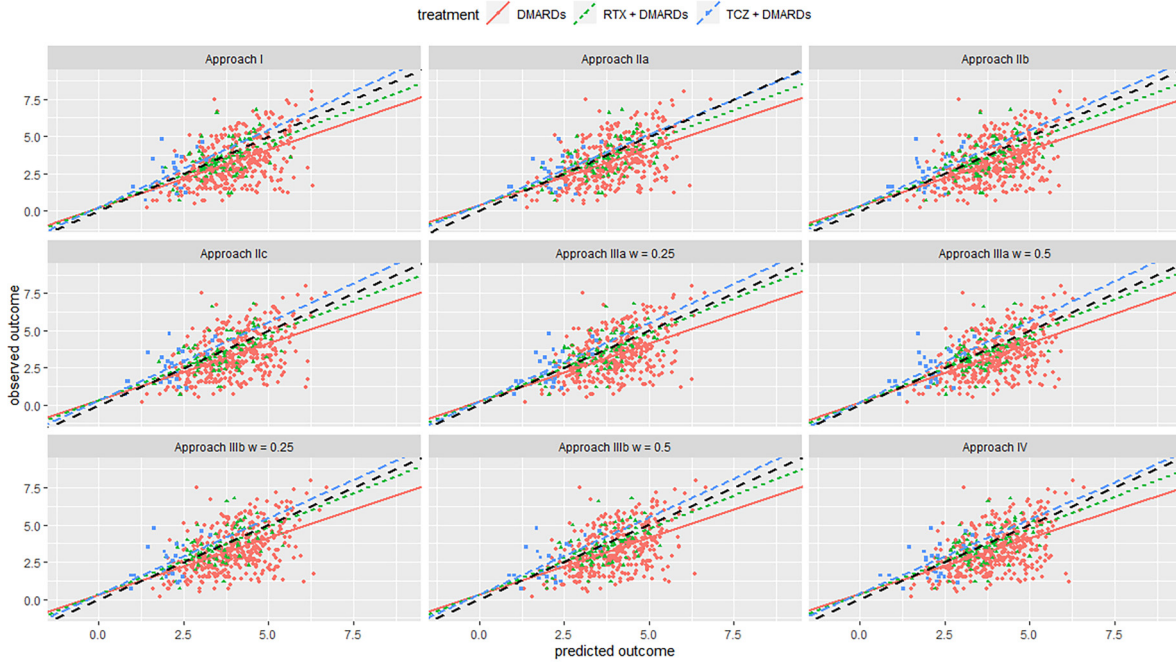


Figure 2. Calibration plot from internal–external validation, for the Swiss registry as the external data set. Black line is line of perfect calibration. Red line is the slope for DMARDs; green line is for RTX + DMARDs; blue line is for TCZ + DMARDs. Each dot represents one patient. Abbreviations: DMARDs: Disease-modifying anti-rheumatic drugs; RTX: rituximab; TCZ: tocilizumab.

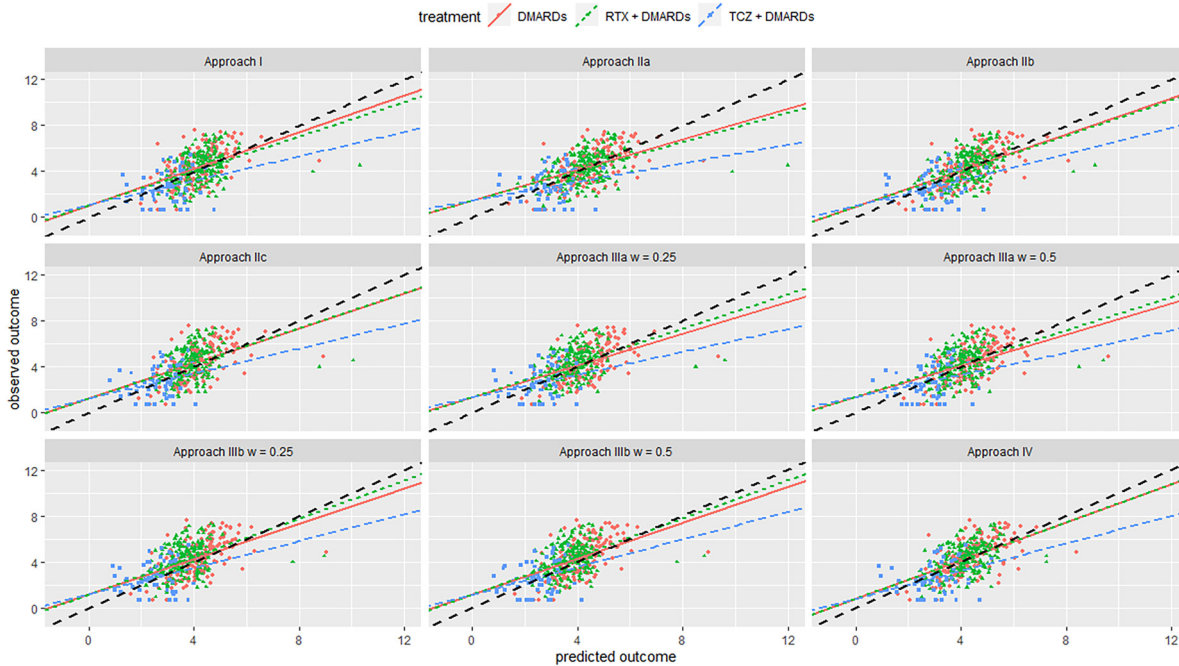


Figure 3. Calibration plot from internal-external validation, for the British registry as the external data set. Black line is line of perfect calibration. Red line is slope for DMARDs; green line is for RTX + DMARDs; the blue line is for TCZ + DMARDs. Each dot represents one patient. Abbreviations: DMARDs: Disease-modifying anti-rheumatic drugs; RTX: rituximab; TCZ: tocilizumab.

Table 4. Internal–external validation results for MSE and bias, for difference approaches.

Left-out data set	Treatment arms	Performance metric	Approach I	Approach IIa	Approach IIb	Approach IIc	Approach IIIa w=0.25	Approach IIIa w=0.5	Approach IIIb w=0.25	Approach IIIb w=0.5	Approach IV		
SCQM	All arms	MSE	1.85	1.74	1.84	1.85	1.87	1.86	1.84	1.84	1.86	1.86	
		Bias	0.51	0.42	0.48	0.48	0.50	0.50	0.51	0.47	0.48	0.50	
	DMARDs	MSE	2.03	1.91	2.03	2.03	2.06	2.05	2.03	2.03	2.03	2.06	
		Bias	0.61	0.52	0.61	0.61	0.65	0.64	0.61	0.61	0.61	0.63	
	RTX +	MSE	1.08	1.07	1.04	1.06	1.04	1.05	1.04	1.04	1.05	1.04	
		Bias	0.20	0.10	0.10	0.08	0.02	0.09	0.09	-0.01	0.07	0.08	
	TCZ +	MSE	1.10	1.02	1.13	1.20	1.12	1.13	1.13	1.15	1.16	1.11	
		Bias	-0.32	-0.29	-0.40	-0.46	-0.37	-0.39	-0.41	-0.41	-0.42	-0.38	
	BSRBR-RA	All arms	MSE	1.67	1.72	1.55	1.71	1.74	1.73	1.79	1.76	1.76	1.52
			Bias	-0.05	0.11	0.02	-0.24	-0.17	-0.09	-0.38	-0.38	-0.35	-0.02
DMARDs		MSE	1.57	1.48	1.41	1.57	1.58	1.60	1.60	1.57	1.57	1.42	
		Bias	-0.19	0.07	0.00	-0.19	0.05	0.11	0.11	-0.19	-0.19	-0.04	
RTX +		MSE	1.67	1.81	1.53	1.80	1.86	1.80	1.80	1.97	1.90	1.45	
		Bias	-0.04	0.08	-0.03	-0.35	-0.39	-0.29	-0.59	-0.59	-0.54	-0.11	
TCZ +		MSE	1.91	2.06	2.03	1.77	1.83	1.84	1.81	1.81	1.81	2.08	
		Bias	0.34	0.29	0.25	-0.05	-0.03	0.03	-0.19	-0.19	-0.18	0.32	
Overall		All arms	MSE	1.76	1.73	1.70	1.78	1.81	1.80	1.82	1.80	1.70	1.70
			Bias	0.25	0.27	0.27	0.14	0.19	0.23	0.08	0.08	0.09	0.26
DMARDs	All arms	MSE	1.90	1.78	1.84	1.90	1.92	1.92	1.90	1.90	1.87	1.87	
		Bias	0.37	0.38	0.43	0.37	0.47	0.48	0.37	0.37	0.37	0.43	
RTX +	All arms	MSE	1.51	1.62	1.40	1.60	1.64	1.60	1.72	1.67	1.34	1.34	
		Bias	0.02	0.09	0.00	-0.23	-0.28	-0.18	-0.44	-0.44	-0.38	-0.06	
TCZ +	All arms	MSE	1.67	1.75	1.76	1.60	1.62	1.63	1.62	1.61	1.79	1.79	
		Bias	0.14	0.12	0.06	-0.17	-0.13	-0.09	-0.25	-0.25	-0.25	0.11	

Abbreviations: MSE: mean squared error; DMARDs: disease-modifying anti-rheumatic drugs; RTX: Rituximab; TCZ: Tocilizumab.

Table 5. Internal-external validation results of the calibration lines and R-squared for different approaches.

Left-out Data set	Performance metric	Approach I	Approach IIa	Approach IIb	Approach IIc	Approach IIIa w = 0.25	Approach IIIa w = 0.5	Approach IIIb w = 0.25	Approach IIIb w = 0.5	Approach IV
SCQM	Calibration slope for outcome	$\eta_0 = 0.25$	$\eta_0 = 0.40$	$\eta_0 = 0.35$	$\eta_0 = 0.34$	$\eta_0 = 0.23$	$\eta_0 = 0.18$	$\eta_0 = 0.33$	$\eta_0 = 0.32$	$\eta_0 = 0.35$
		$\eta_1 = 0.78$	$\eta_1 = 0.76$	$\eta_1 = 0.76$	$\eta_1 = 0.76$	$\eta_1 = 0.78$	$\eta_1 = 0.79$	$\eta_1 = 0.76$	$\eta_1 = 0.76$	$\eta_1 = 0.75$
		$\eta_2 = 0.88$	$\eta_2 = 0.86$	$\eta_2 = 0.88$	$\eta_2 = 0.88$	$\eta_2 = 0.93$	$\eta_2 = 0.92$	$\eta_2 = 0.91$	$\eta_2 = 0.89$	$\eta_2 = 0.88$
		$\eta_3 = 1.04$	$\eta_3 = 0.96$	$\eta_3 = 1.02$	$\eta_3 = 1.05$	$\eta_3 = 1.05$	$\eta_3 = 1.08$	$\eta_3 = 1.03$	$\eta_3 = 1.04$	$\eta_3 = 1.03$
		$\kappa_0 = 0.22$	$\kappa_0 = 0.35$	$\kappa_0 = 0.27$	$\kappa_0 = 0.24$	$\kappa_0 = 0.14$	$\kappa_0 = 0.09$	$\kappa_0 = 0.23$	$\kappa_0 = 0.23$	$\kappa_0 = 0.25$
		$\kappa_1 = 0.79$	$\kappa_1 = 0.78$	$\kappa_1 = 0.78$	$\kappa_1 = 0.78$	$\kappa_1 = 0.80$	$\kappa_1 = 0.82$	$\kappa_1 = 0.79$	$\kappa_1 = 0.79$	$\kappa_1 = 0.78$
	Calibration slope for benefit	$\kappa_2 = 0.23$	$\kappa_2 = 0.29$	$\kappa_2 = 0.23$	$\kappa_2 = 0.20$	$\kappa_2 = 0.19$	$\kappa_2 = 0.21$	$\kappa_2 = 0.19$	$\kappa_2 = 0.20$	$\kappa_2 = 0.20$
		$\kappa_3 = 0.45$	$\kappa_3 = 0.53$	$\kappa_3 = 0.46$	$\kappa_3 = 0.43$	$\kappa_3 = 0.44$	$\kappa_3 = 0.44$	$\kappa_3 = 0.44$	$\kappa_3 = 0.44$	$\kappa_3 = 0.44$
		0.09	0.14	0.09	0.09	0.08	0.09	0.09	0.09	0.08
		$\eta_0 = 1.07$	$\eta_0 = 1.47$	$\eta_0 = 0.97$	$\eta_0 = 1.23$	$\eta_0 = 1.35$	$\eta_0 = 1.36$	$\eta_0 = 1.21$	$\eta_0 = 1.15$	$\eta_0 = 0.81$
		$\eta_1 = 0.80$	$\eta_1 = 0.66$	$\eta_1 = 0.79$	$\eta_1 = 0.76$	$\eta_1 = 0.69$	$\eta_1 = 0.68$	$\eta_1 = 0.77$	$\eta_1 = 0.78$	$\eta_1 = 0.83$
		$\eta_2 = 0.75$	$\eta_2 = 0.64$	$\eta_2 = 0.77$	$\eta_2 = 0.77$	$\eta_2 = 0.75$	$\eta_2 = 0.72$	$\eta_2 = 0.82$	$\eta_2 = 0.83$	$\eta_2 = 0.83$
BSRBR-RA	Calibration slope for outcome	$\eta_3 = 0.53$	$\eta_3 = 0.40$	$\eta_3 = 0.56$	$\eta_3 = 0.55$	$\eta_3 = 0.50$	$\eta_3 = 0.48$	$\eta_3 = 0.58$	$\eta_3 = 0.60$	$\eta_3 = 0.60$
		$\kappa_0 = 1.11$	$\kappa_0 = 1.10$	$\kappa_0 = 0.87$	$\kappa_0 = 1.02$	$\kappa_0 = 1.17$	$\kappa_0 = 1.23$	$\kappa_0 = 1.00$	$\kappa_0 = 0.98$	$\kappa_0 = 0.84$
		$\kappa_1 = 0.75$	$\kappa_1 = 0.70$	$\kappa_1 = 0.78$	$\kappa_1 = 0.78$	$\kappa_1 = 0.72$	$\kappa_1 = 0.69$	$\kappa_1 = 0.80$	$\kappa_1 = 0.81$	$\kappa_1 = 0.81$
		$\kappa_2 = 0.59$	$\kappa_2 = 0.36$	$\kappa_2 = 0.59$	$\kappa_2 = 0.47$	$\kappa_2 = 0.38$	$\kappa_2 = 0.39$	$\kappa_2 = 0.45$	$\kappa_2 = 0.48$	$\kappa_2 = 0.75$
		$\kappa_3 = 1.32$	$\kappa_3 = 0.98$	$\kappa_3 = 1.08$	$\kappa_3 = 1.03$	$\kappa_3 = 0.95$	$\kappa_3 = 0.95$	$\kappa_3 = 0.98$	$\kappa_3 = 0.99$	$\kappa_3 = 1.20$
		0.27	0.25	0.32	0.25	0.23	0.24	0.21	0.23	0.33
	Calibration slope for benefit	$\eta_0 = 1.11$	$\eta_0 = 1.10$	$\eta_0 = 1.08$	$\eta_0 = 1.03$	$\eta_0 = 1.03$	$\eta_0 = 1.03$	$\eta_0 = 1.03$	$\eta_0 = 1.03$	$\eta_0 = 1.03$
		$\eta_1 = 0.80$	$\eta_1 = 0.66$	$\eta_1 = 0.79$	$\eta_1 = 0.76$	$\eta_1 = 0.69$	$\eta_1 = 0.68$	$\eta_1 = 0.77$	$\eta_1 = 0.78$	$\eta_1 = 0.83$
		$\eta_2 = 0.75$	$\eta_2 = 0.64$	$\eta_2 = 0.77$	$\eta_2 = 0.77$	$\eta_2 = 0.75$	$\eta_2 = 0.72$	$\eta_2 = 0.82$	$\eta_2 = 0.83$	$\eta_2 = 0.83$
		$\eta_3 = 1.04$	$\eta_3 = 0.96$	$\eta_3 = 1.02$	$\eta_3 = 1.05$	$\eta_3 = 1.05$	$\eta_3 = 1.08$	$\eta_3 = 1.03$	$\eta_3 = 1.04$	$\eta_3 = 1.03$
		$\kappa_0 = 0.22$	$\kappa_0 = 0.35$	$\kappa_0 = 0.27$	$\kappa_0 = 0.24$	$\kappa_0 = 0.14$	$\kappa_0 = 0.09$	$\kappa_0 = 0.23$	$\kappa_0 = 0.23$	$\kappa_0 = 0.25$
		$\kappa_1 = 0.79$	$\kappa_1 = 0.78$	$\kappa_1 = 0.78$	$\kappa_1 = 0.78$	$\kappa_1 = 0.80$	$\kappa_1 = 0.82$	$\kappa_1 = 0.79$	$\kappa_1 = 0.79$	$\kappa_1 = 0.78$
R squared	$\kappa_2 = 0.23$	$\kappa_2 = 0.29$	$\kappa_2 = 0.23$	$\kappa_2 = 0.20$	$\kappa_2 = 0.19$	$\kappa_2 = 0.21$	$\kappa_2 = 0.19$	$\kappa_2 = 0.20$	$\kappa_2 = 0.20$	
	$\kappa_3 = 0.45$	$\kappa_3 = 0.53$	$\kappa_3 = 0.46$	$\kappa_3 = 0.43$	$\kappa_3 = 0.44$	$\kappa_3 = 0.44$	$\kappa_3 = 0.44$	$\kappa_3 = 0.44$	$\kappa_3 = 0.44$	
	0.09	0.14	0.09	0.09	0.08	0.09	0.09	0.09	0.08	
	$\eta_0 = 1.07$	$\eta_0 = 1.47$	$\eta_0 = 0.97$	$\eta_0 = 1.23$	$\eta_0 = 1.35$	$\eta_0 = 1.36$	$\eta_0 = 1.21$	$\eta_0 = 1.15$	$\eta_0 = 0.81$	
	$\eta_1 = 0.80$	$\eta_1 = 0.66$	$\eta_1 = 0.79$	$\eta_1 = 0.76$	$\eta_1 = 0.69$	$\eta_1 = 0.68$	$\eta_1 = 0.77$	$\eta_1 = 0.78$	$\eta_1 = 0.83$	
	$\eta_2 = 0.75$	$\eta_2 = 0.64$	$\eta_2 = 0.77$	$\eta_2 = 0.77$	$\eta_2 = 0.75$	$\eta_2 = 0.72$	$\eta_2 = 0.82$	$\eta_2 = 0.83$	$\eta_2 = 0.83$	

These follow from linear regressions of the observed vs. predicted, as noted in equations (13) and (14). Subscript 0 refers to the intercept. 1 refers to DMARDs, 2 refers to RTX + DMARDs, and 3 refers to TCZ + DMARDs. Abbreviations: DMARDs: disease-modifying anti-rheumatic drugs; RTX: rituximab; TCZ: tocilizumab.

the sample size was big, the outcome was (approximately) continuous, and we only used few predictors. Finally, results did not materially change when using different prior distributions for λ .

Discussion

This paper presents a general framework for developing methods to predict real-world outcomes for a range of different treatment options. We focussed on continuous outcomes and proposed various prediction modelling approaches and methods for assessing their performance. Our models were based on two-stage IPD NMA and utilized both randomized and observational data.

We used a data set of patients with RA obtained from three RCTs and two registries to illustrate our methods. For this example, we developed six meta-analytical models and a simpler model (i.e. Approach I). The latter only utilized data from a single registry at a time. After fitting each model, we assessed its internal and internal–external performance while adjusting for optimism. In these validations, we compared observed outcomes versus predictions for all registry patients and calculated bias, MSE, and R-squared. We also fitted calibration lines, both for the overall outcome as well as for treatment benefit.

For our example, internal validation showed that Approach I was among the best performing approaches for both the Swiss and the British registry. Thus, for this particular example, the incorporation of RCT data and advanced meta-analytical modelling brought small benefit in terms of predictive ability for patients found in these two registries. Conversely, the internal–external cross-validation procedure identified Approaches IIb and IV as the best-performing ones. This suggested that these two might be the best for patients found in a completely new setting among all developed models. However, we should keep in mind that to establish the generalizability of a model, we would need data from multiple real-world settings to be available. In case of important heterogeneity (i.e. when there are great differences across different real-world settings), it will be tough to make predictions about patients found in new settings. This is one general limitation of all proposed methods.

Another important limitation of our approach is the need for IPD from multiple sources. IPD is generally hard to obtain at the MA level. However, new initiatives have been recently launched, such as YODA (<https://yoda.yale.edu/>), Vivli (<https://vivli.org/>), and Clinical Study Data Request (<https://www.clinicalstudydatarequest.com/>), aiming to promote large-scale IPD sharing. Moreover, our approach requires a connected network of interventions. In practical applications, the available NRSs may not include all interventions of interest (despite forming a closed network). In such cases, although we can still use some of the described methods to make predictions for outcomes under the missing treatments, there is no way to assess the goodness of these predictions for these particular settings. Another limitation is that we only discussed the

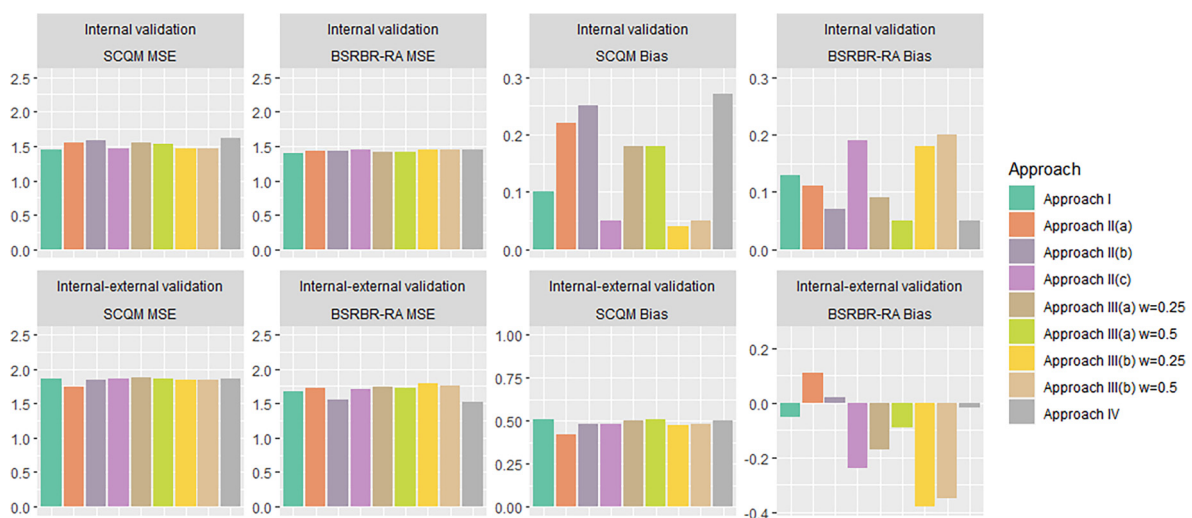


Figure 4. Bar plot summarizing MSE and bias calculated through an internal (top row) and internal–external (bottom row) validation, for the Swiss and British registry (SCQM and BSRBR-RA respectively). For internal–external validation, the labelled study is used as the target data set. Abbreviations: MSE: Mean squared error; DMARDs: Disease-modifying anti-rheumatic drugs; RTX: rituximab; TCZ: tocilizumab.

case of comparative studies, i.e. studies that included more than one treatment. A generalization of the methods to include single-arm studies might be of interest in a follow-up project.

An additional limitation is that the proposed linear model (with only two-way interactions between treatment and covariate) might be too simple to predict accurately in the real-world. If a non-linear relationship is suspected, we could consider models that include fractional polynomials or cubic splines. Alternative approaches include tree-based methods.^{40–42} However, we did not explore them in detail.

In Approach III, we used fixed weighting factors w_j . We arbitrarily used values of 0.25 and 0.5 for these factors. Sensitivity analysis showed that using different values did not alter much the performance of the model. Alternatively, we could have used flexible weighting, e.g. by assigning a prior distribution to w_j , treating the weight as a random variable.¹⁵ This is similar to the ‘power prior’ approach.^{43,44} We fitted a model that used a modification of the power prior approach, called normalized power prior^{45,46} in the RA data set. In this model, the parameter estimates from NRS are down-weighted according to whether they agree with the RCT estimates. In the case of a large disagreement, there is more down-weighting. We found that the weighting factors decreased all the way down to zero for the RA data set, negating all information from the NRS. For large sample data sets the RCTs and NRS are likely to disagree (i.e. large discrepancies between the mean and variance of RCTs and NRS), which will lead to pronounced down-weighting.⁴⁷ Thus, we did not show models based on the power prior in this paper; however, this might be an interesting area for future development. Also, we could assign weights to different model parameters, e.g. the intercept term, if deemed appropriate; we did not pursue this further in this paper. Another limitation of our methods is that it may often be difficult to use calibration measures (i.e. calibration intercept and slope for predicted outcome and predicted benefit) to compare the performance between competing models. It is possible that some models yield accurate outcome predictions (i.e. prognosis) but not of treatment benefit, and vice versa. Selecting an appropriate model will then strongly depend on the context in which the model should be used.

Moreover, we could have explored approaches where penalization also applies to prognostic factors and not only to effect modifiers, and also other penalization methods. Also, we only discussed two-stage meta-analytical approaches; however, making these methods one-stage is straightforward. Another interesting idea suggested by one of our anonymous reviewers would be to include study design as yet another predictor in the second-stage model. However, it would be difficult to implement in our data set, which only includes two NRS; we leave this idea for future research.

In this paper, we used a simple regression adjustment for estimating treatment effects from the observational studies. Alternative methods such as propensity score matching⁴⁸ or inverse probability of treatment weighting (IPTW) could be used when the assumptions behind the regression adjustment seem implausible. We explored the use of IPTW and preliminary results showed a similar performance to the models based on regression adjustment. Thus, we decided to leave this for future work. Note, however, that no method for causal inference guarantees unbiased estimation of treatment effects from the observational studies. In our framework, when these estimates are biased, models that use this information (such as Approach IIa or IIb) are expected to perform worse, and thus not be selected at the end as the final model. Also, note that all methods for causal inference are limited by unmeasured confounding.

Furthermore, in practical applications, we often face the problem of ‘systematically missing’ predictors, i.e. when a predictor is missing for all individuals within particular studies in an IPD MA. There have been methods proposed for imputing such predictors, based on the missing at random assumption.⁴⁹ There is also an R package `micemd`,⁵⁰ which implements some of these imputation methods. This however requires all data to be on the same server; in our experience, this cannot always be achieved, especially when data from different studies are owned by different for-profit companies. In such cases, researchers may only get access to data via the companies’ private servers, without being able to merge data sets from different studies.

Finally, we note that our methods need to be extended to cover the case of binary and time-to-event outcomes. This might be challenging, especially when assessing the predictive performance of different approaches using calibration for treatment benefit. For example, the so-called c-statistic for benefit was recently proposed⁵¹ and might be used to this end; however, future research is needed to investigate this.

To summarize, this is the first paper to propose and compare two-stage meta-analytical IPD models for predicting the real-world effectiveness of interventions to the best of our knowledge. The gain in predictive performance from combining RCTs and NRS was modest in our clinical example. Nevertheless, the illustration of different modelling approaches and the considerations regarding different cross-validation methods that we provide may be valuable to inform future studies aiming to predict real-world outcomes of competing interventions.

Acknowledgements

A list of rheumatology offices and hospitals that are contributing to the SCQM registries can be found on www.scqm.ch/institutions. The SCQM is financially supported by pharmaceutical industries and donors. A list of financial supporters can be found on www.scqm.ch/sponsors. The author(s) would like to thank the British Society for Rheumatology for use of this data. The BSR Biologics Register in

Rheumatoid Arthritis (BSRBR-RA) is a project owned and run by the BSR on behalf of its members. BSR would like to thank its funders for their support, currently AbbVie, Amgen, Celltrion, Eli Lilly, Pfizer, Samsung, Sandoz and Sanofi, and in the past Merck, Roche, Swedish Orphan Biovitrum (SOBI) and UCB. This income finances a wholly separate contract between BSR and The University of Manchester to host the BSRBR-RA. All decisions concerning analyses, interpretation and publication are made autonomously of any industrial contribution. The BSRBR-RA would like to gratefully acknowledge the support of BSR members and the National Institute for Health Research, through the Comprehensive Local Research Networks at participating centres; The authors would like to thank F. Hoffmann-La Roche and Clinical Study Data Request (<https://clinicalstudydatarequest.com/>) for providing the trial data.

Author Note

Orestis Efthimiou, Institute of Primary Health Care (BIHAM), University of Bern, Bern, Switzerland.






Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: MS and OE were supported by the Swiss National Science Foundation (Ambizione grant number 180083). ME was supported by special project funding (grant 17841) from the Swiss National Science Foundation. SB was supported by the Medical Research Council (grant no. MR/R025223/1). TD acknowledges financial support from the Netherlands Organization for Health Research and Development (grant 91617050), and the European Union's Horizon 2020 Research and Innovation Programme under ReCoDID Grant Agreement no. 825746. This project has received funding from the European Union's Horizon 2020 research and innovation programme under ReCoDID grant agreement no. 825746.

ORCID iDs

Michael Seo  <https://orcid.org/0000-0002-5229-590X>
 Thomas PA Debray  <https://orcid.org/0000-0002-1790-2719>
 Yann Ruffieux  <https://orcid.org/0000-0002-0891-2448>
 Sandro Gsteiger  <https://orcid.org/0000-0001-9775-9088>
 Orestis Efthimiou  <https://orcid.org/0000-0002-0955-7572>

Supplemental Material

Supplemental material for this article is available online.

References

1. Evans D. Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. *J Clin Nurs* 2003; **12**: 77–84.
2. Revicki DA and Frank L. Pharmacoeconomic evaluation in the real world. Effectiveness versus efficacy studies. *Pharmacoeconomics* 1999; **15**: 423–434.
3. Eichler H-G, Abadie E, Breckenridge A, et al. Bridging the efficacy-effectiveness gap: a regulator's perspective on addressing variability of drug response. *Nat Rev Drug Discov* 2011; **10**: 495–506.
4. Nordon C, Karcher H, Groenwold RHH, et al. The 'efficacy-effectiveness gap': historical background and current conceptualization. *Value Health* 2016; **19**: 75–81.
5. Reeves BC, Higgins JPT, Ramsay C, et al. An introduction to methodological issues when including non-randomised studies in systematic reviews on the effects of interventions. *Res Synth Methods* 2013; **4**: 1–11.
6. Schünemann HJ, Tugwell P, Reeves BC, et al. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. *Res Synth Methods* 2013; **4**: 49–62.
7. Concato J, Shah N and Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000; **342**: 1887–1892.
8. Sarri G, Patomo E, Yuan H, et al. Framework for the synthesis of non-randomised studies and randomised controlled trials: a guideline on conducting a systematic review and meta-analysis for healthcare decision making. *BMJ EBM* 2020; **27**: 109–119.
9. Turner RM, Spiegelhalter DJ, Smith GCS, et al. Bias modelling in evidence synthesis. *J R Stat Soc Ser A Stat Soc* 2009; **172**: 21–47.
10. Ferguson J, Alvarez-Iglesias A, Newell J, et al. Joint incorporation of randomised and observational evidence in estimating treatment effects. *Stat Methods Med Res* 2019; **28**: 235–247.
11. Verde PE and Ohmann C. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. *Res Synth Methods* 2015; **6**: 45–62.

12. Verde PE. A bias-corrected meta-analysis model for combining, studies of different types and quality. *Biometrical Journal* 2020; **63**: 406–422. Epub ahead of print 30 September 2020. DOI: 10.1002/bimj.201900376.
13. Cameron C, Fireman B, Hutton B, et al. 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. *Syst Rev* 2015; **4**: 147–147.
14. Schmitz S, Adams R and Walsh C. Incorporating data from various trial designs into a mixed treatment comparison model. *Stat Med* 2013; **32**: 2935–2949.
15. Efthimiou O, Mavridis D, Debray TPA, et al. Combining randomized and non-randomized evidence in network meta-analysis. *Stat Med* 2017; **36**: 1210–1226.
16. Barry MJ and Edgman-Levitan S. Shared decision making — the pinnacle of patient-centered care. *N Engl J Med* 2012; **366**: 780–781.
17. Didden E-M, Ruffieux Y, Hummel N, et al. Prediction of real-world drug effectiveness prelaunch: case study in rheumatoid arthritis. *Med Decis Making* 2018; **38**: 719–729.
18. Efthimiou O, Debray TPA, Valkenhoef G van, et al. Getreal in network meta-analysis: a review of the methodology. *Res Synth Methods* 2016; **7**: 236–263.
19. Debray TPA, Moons KGM, van Valkenhoef G, et al. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. *Res Synth Methods* 2015; **6**: 293–309.
20. Grassi W, De Angelis R, Lamanna G, et al. The clinical features of rheumatoid arthritis. *Eur J Radiol* 1998; **27**: S18–S24.
21. Fransen J, Stucki G and van Riel PLCM. Rheumatoid arthritis measures: disease activity score (DAS), disease activity score-28 (DAS28), rapid assessment of disease activity in rheumatology (RADAR), and rheumatoid arthritis disease activity Index (RADAI). *Arthritis Care Res (Hoboken)* 2003; **49**: S214–S224.
22. Dennison EM, Packham J and Hyrich K. The BSRBR-RA at 15 years. *Rheumatology (Oxford)* 2016; **55**: 2093–2095.
23. Langenegger T, Fransen J, Forster A, et al. Klinisches Qualitätsmanagement bei der rheumatoiden Arthritis. *Zeitschrift für Rheumatologie* 2001; **60**: 333–341.
24. Steyerberg EW. *Clinical prediction models*. 2nd Ed. New York: Springer, 2019.
25. Hastie T, Tibshirani R and Friedman J. *The elements of statistical learning : data mining, inference, and prediction*. New York: Springer, 2001: 43–99.
26. Seo M, White IR, Furukawa TA, et al. Comparing methods for estimating patient-specific treatment effects in individual patient data meta-analysis. *Stat Med* 2021; **40**: 1553–1573.
27. Park T and Casella G. The Bayesian lasso. *J Am Stat Assoc* 2008; **103**: 681–686.
28. O'Hara RB and Sillanpää MJ. A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal* 2009; **4**: 85–117.
29. Lykou A and Ntzoufras I. On Bayesian lasso variable selection and the specification of the shrinkage parameter. *Stat Comput* 2013; **23**: 361–390.
30. Debray TP, Schuit E, Efthimiou O, et al. An overview of methods for network meta-analysis using individual participant data: when do benefits arise? *Stat Methods Med Res* 2018; **27**: 1351–1364.
31. Turner RM, Jackson D, Wei Y, et al. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Stat Med* 2015; **34**: 984–998.
32. Steyerberg EW and Harrell FE Jr. Prediction models need appropriate internal, internal–external, and external validation. *J Clin Epidemiol* 2016; **69**: 245–247.
33. Debray TPA, Moons KGM, Ahmed I, et al. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med* 2013; **32**: 3158–3180.
34. Gelman A, Carlin JB, Stern HS, et al. *Bayesian Data analysis*. 2nd ed. New York: Chapman and Hall/CRC, 2004.
35. van Buuren S and Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw* 2011; **45**: 67.
36. Zhou X and Reiter JP. A note on Bayesian inference after multiple imputation. *Am Stat* 2010; **64**: 159–163.
37. Core Team R. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, <https://www.R-project.org/> (2018 accessed 12 October 2021).
38. Plummer M. Rjags: bayesian graphical models using MCMC, <https://CRAN.R-project.org/package=rjags> (2019).
39. Gelman A and Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci* 1992; **7**: 457–472.
40. Li X, Dusseldorp E and Meulman JJ. A flexible approach to identify interaction effects between moderators in meta-analysis. *Res Synth Methods* 2019; **10**: 134–152.
41. Fokkema M, Smits N, Zeileis A, et al. Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behav Res Methods* 2018; **50**: 2016–2034.
42. Seibold H, Zeileis A and Hothorn T. Individual treatment effect prediction for amyotrophic lateral sclerosis patients. *Stat Methods Med Res* 2018; **27**: 3104–3125.
43. Ibrahim JG and Chen M-H. Power prior distributions for regression models. *Statist Sci* 2000; **15**: 46–60.
44. Ibrahim JG, Chen M-H, Gwon Y, et al. The power prior: theory and applications. *Stat Med* 2015; **34**: 3724–3749.
45. Duan Y, Ye K and Smith EP. Evaluating water quality using power priors to incorporate historical information. *Environmetrics* 2006; **17**: 95–106.
46. Neuenschwander B, Branson M and Spiegelhalter DJ. A note on the power prior. *Stat Med* 2009; **28**: 3562–3566.

47. Neelon B, James A and Malley O. Bayesian Analysis using power priors with application to pediatric quality of care. *J Biom Biostat* 2010; **1**: 1–9. Epub ahead of print 1 January 2010. DOI: 10.4172/2155-6180.1000103.
48. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
49. Audigier V, White I, Jolani S, et al. Multiple imputation for multilevel data with continuous and binary variables. *Stat Sci* 2018; **33**: 160–183. Epub ahead of print 2018. DOI: 10.1214/18-STS646.
50. Audigier V, Resche-Rigon M. *micemd: multiple Imputation by Chained Equations with Multilevel Data*, <https://CRAN.R-project.org/package=micemd> (2018).
51. van Klaveren D, Steyerberg EW, Serruys PW, et al. The proposed ‘concordance-statistic for benefit’ provided a useful metric when modeling heterogeneous treatment effects. *J Clin Epidemiol* 2018; **94**: 59–68.