

This is a repository copy of *Combining randomised and non-randomized data to predict heterogeneous effects of competing treatments*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/210214/>

Version: Published Version

---

**Article:**

Chalkou, Konstantina, Hamza, Tasnim, Benkert, Pascal et al. (8 more authors) (2024) Combining randomised and non-randomized data to predict heterogeneous effects of competing treatments. *Research Synthesis Methods*. ISSN 1759-2887

<https://doi.org/10.1002/jrsm.1717>

---

**Reuse**







This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

# Combining randomized and non-randomized data to predict heterogeneous effects of competing treatments

Konstantina Chalkou<sup>1,2,3</sup>  | Tasnim Hamza<sup>1,2</sup>  | Pascal Benkert<sup>4</sup>  |  
 Jens Kuhle<sup>5,6,7,8</sup> | Chiara Zecca<sup>9,10</sup>  | Gabrielle Simoneau<sup>11</sup> |  
 Fabio Pellegrini<sup>12</sup> | Andrea Manca<sup>13</sup>  | Matthias Egger<sup>1,14</sup>  | Georgia Salanti<sup>1</sup>

<sup>1</sup>Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

<sup>2</sup>Graduate School for Health Sciences, University of Bern, Bern, Switzerland

<sup>3</sup>Department of Clinical Research, University of Bern, Bern, Switzerland

<sup>4</sup>Department of Clinical Research, University Hospital Basel, University of Basel, Basel, Switzerland

<sup>5</sup>Multiple Sclerosis Centre, Neurologic Clinic and Policlinic, Department of Head, Spine and Neuromedicine, University Hospital Basel, University of Basel, Basel, Switzerland

<sup>6</sup>Multiple Sclerosis Centre, Neurologic Clinic and Policlinic, Department of Biomedicine, University Hospital Basel, University of Basel, Basel, Switzerland

<sup>7</sup>Multiple Sclerosis Centre, Neurologic Clinic and Policlinic, Department of Clinical Research, University Hospital Basel, University of Basel, Basel, Switzerland

<sup>8</sup>Research Center for Clinical Neuroimmunology and Neuroscience (RC2NB), University Hospital, University of Basel, Basel, Switzerland

<sup>9</sup>Multiple Sclerosis Center, Neurocenter of Southern Switzerland, EOC, Lugano, Switzerland

<sup>10</sup>Faculty of Biomedical Sciences, Università della Svizzera Italiana, Lugano, Switzerland

<sup>11</sup>Biogen Canada, Toronto, Ontario, Canada

<sup>12</sup>Biogen Digital Health, Biogen Spain, Madrid, Spain

<sup>13</sup>Centre for Health Economics, University of York, York, UK

<sup>14</sup>Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

## Correspondence

Konstantina Chalkou, Institute of Social and Preventive Medicine, University of Bern, Mittelstrasse 43, 3012 Bern, Switzerland.

Email: [kontantina.chalkou@unibe.ch](mailto:kontantina.chalkou@unibe.ch)

## Funding information

European Union's Horizon 2020 research and innovation programme, Grant/Award Number: No 825162; National Science Foundation, Grant/Award Number: 189498

## Abstract

Some patients benefit from a treatment while others may do so less or do not benefit at all. We have previously developed a two-stage network meta-regression prediction model that synthesized randomized trials and evaluates how treatment effects vary across patient characteristics. In this article, we extended this model to combine different sources of types in different formats: aggregate data (AD) and individual participant data (IPD) from randomized and non-randomized evidence. In the first stage, a prognostic model is developed to predict the baseline risk of the outcome using a large cohort study. In the second stage, we recalibrated this prognostic model to improve our predictions for patients enrolled in randomized trials. In the third stage, we used the baseline risk as effect modifier in a network meta-regression model combining AD, IPD randomized clinical trial to estimate heterogeneous treatment effects.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *Research Synthesis Methods* published by John Wiley & Sons Ltd.

We illustrated the approach in the re-analysis of a network of studies comparing three drugs for relapsing–remitting multiple sclerosis. Several patient characteristics influence the baseline risk of relapse, which in turn modifies the effect of the drugs. The proposed model makes personalized predictions for health outcomes under several treatment options and encompasses all relevant randomized and non-randomized evidence.

#### KEYWORDS

combination of data sources, network meta-analysis, prediction model

#### Highlights

##### What is already known?

- Recently, a two-stage model which allows for individualized treatment effects predictions between several competing treatments was developed, using individual participant data from a network of randomized clinical trials.

##### What is new?

- We extend this model by combining several data sources, such as observational studies on the top of randomized clinical trials and we show how to incorporate aggregate data in the analysis.

##### Potential impact for *Research Synthesis Methods* readers outside authors' field

- Readers will be able to reproduce the suggested model, in any clinical area, to make individualized predictions of several competing treatments based on network meta-analysis results, while combining several data sources; the methods are described in detail and the codes used for the illustrative example are publicly available.

## 1 | INTRODUCTION

Applications of network meta-analysis to health care questions typically report population-average results about the effects of competing treatments.<sup>1,2</sup> The applicability of such results is limited for decision-making purposes, as some patients might benefit greatly from a treatment while others may do so less or do not benefit at all. Network meta-regression models of studies with individual participant data (IPD) can be used to estimate such heterogeneous treatments effects, should preferably account for all relevant individual's baseline characteristics simultaneously, and indicate the preferable treatment for each patient.<sup>3–5</sup> The most commonly used methods for individualized predictions are the effect modification and the risk modeling approaches.<sup>3,6–10</sup>

The *effect modification approach* is a regression model with multiple variables and interaction terms between them and the treatment.<sup>11</sup> The effect modification

approach is a flexible method for predicting individualized treatment effects but presents some practical difficulties.<sup>8,11</sup> First, it is prone to overfitting because often a large number of model parameters must be estimated from a small or insufficient sample size.<sup>12–14</sup> Although penalization approaches could potentially alleviate the risk of overfitting, research on penalization in (network) meta-regression models is still at an experimental phase.<sup>9</sup> Risk modeling approaches have been developed as a solution to these shortcomings.

The *risk modeling approach* is a two-stage method to estimate heterogeneous treatment effects within a trial. The approach assumes that the risk of the outcome estimated at baseline (often a proxy for the severity of the condition, the presence of comorbidities, etc.) could moderate the treatment effects.<sup>3–5,8,11,15</sup> In the *first stage*, the outcome risk for each patient is predicted according to their characteristics at baseline. In the *second stage*, the interaction between the baseline risk and the treatment

effect is estimated.<sup>3,8,16–20</sup> The same trial data (internal risk modeling) or different datasets (external risk modeling) can be used at each stage.<sup>3–5,11</sup> The risk modeling approach can be thought of as a parameter reduction method, which reduces the risk of overfitting which is one of the most important problems when dealing with a large number of treatment–covariate interactions. Risk modeling outperforms the effect modification method in terms of dimensionality, power, and when there is limited prior knowledge about the role of covariates, while taking advantage of well-established penalization and variable selection methods in multivariable prognostic models.<sup>8</sup> Analysis of randomized clinical trials (RCTs) using risk modeling has been successfully applied in various clinical areas to estimate heterogeneous treatment effects.<sup>8,9,11,18</sup> The internal risk modeling approach was recently extended into a network meta-regression model of RCTs with IPD.<sup>21</sup>

This work is supported and funded by the HTx. The HTx is a Horizon 2020 project supported by the European Union lasting for 5 years from January 2019. The main aim of HTx is to create a framework for the Next Generation Health Technology Assessment to support patient-centered, societally oriented, real-time decision-making on access to and reimbursement for health technologies throughout Europe. To this end, the aim of this paper is to extend the previously developed two-stage risk modeling approach, using a fusion of evidence synthesis and prediction methodology.<sup>21</sup> Observational rather than randomized studies are arguably more suitable to develop a prognostic model,<sup>22</sup> while RCTs are more suitable to estimate unbiased treatment effects. Consequently, we extend the approach so that different sources of data are employed in the different stages of the risk modeling approach. The differences in populations in observational studies and RCTs need to be accounted for and we suggest re-calibration techniques for this purpose. To increase power and precision of the estimated treatment effects, we also suggest that studies that report only aggregated data (AD) could be included in the approach.<sup>23</sup> We implemented the network meta-regression model in a Bayesian framework, and we used it to predict the probability of experiencing at least one relapse within the next 2 years for three drugs and placebo in patients with relapsing–remitting multiple sclerosis (RRMS).

## 2 | MOTIVATING EXAMPLE AND DATA

Multiple sclerosis is an immune-mediated disease of the central nervous system with various subtypes. Its most common subtype is RRMS.<sup>24</sup> Patients with RRMS present

with acute or subacute symptoms (relapses) followed by periods of complete or incomplete recovery (remissions).<sup>25</sup> Reduction in relapse rates has been commonly used as the primary efficacy endpoint in phase III RCTs leading to market approval of drugs.<sup>24</sup> Some of the drugs, like natalizumab, are associated with rare but serious side effects, while others, like dimethyl fumarate, are considered to be safer options.<sup>26,27</sup>

We illustrate the methods in datasets including patients with confirmed RRMS. Table 1 presents the outcome and the patients' baseline characteristics for the included datasets. The outcome of interest is at least on relapse within 2 years from baseline (yes or no).

### 2.1 | Observational evidence

We included 935 patients enrolled in the Swiss Multiple Sclerosis Cohort (SMSC).<sup>28</sup> Each patient contributed one, two, or three cycles of 2 years of follow-up. The beginning of each 2-year follow-up cycle is assumed as time zero, which corresponds to the moment when a decision needs to be made regarding the initiation or revision of treatment, after “baseline” demographic (such as age, gender) and disease activity (e.g., Expanded Disability Status Scale [EDSS]) data are collected. We included 1752 patient cycles in total.

### 2.2 | Randomized evidence

We had access to IPD from three phase-III RCTs with 3590 patients assigned to placebo, natalizumab, dimethyl fumarate, or glatiramer acetate.<sup>29–31</sup> We included a subset of 2150 patients with complete covariate and outcome information, assuming that any missingness does not depend on the risk of relapsing.

## 3 | METHODS

We proposed a three-stage model. In the first stage, we built a prognostic model using the SMSC. In the second stage, we recalibrated the model to estimate the baseline risk in patients enrolled in RCTs. In the third stage, we estimated heterogeneous treatment effects from a network meta-regression model that synthesizes AD with IPD from RCTs and includes the baseline risk as a prognostic factor and effect modifier.

All our analyses were done in R<sup>32</sup> version 3.6.2 and in JAGS<sup>33</sup> (called through R). The code can be found in the GitHub repository: [https://github.com/esm-isp-unibe-ch/ThreeStageModel\\_RRMS](https://github.com/esm-isp-unibe-ch/ThreeStageModel_RRMS).

TABLE 1 Treatment, sample size, outcome, baseline characteristics, and baseline risk (stage 2) of patients in the included datasets.

Study (type of data)	Treatment	Number of patients	Number of patients experiencing relapse at 2 years (%)	Mean age (sd)	Number of females (%)	Mean baseline EDSS score (sd)	Mean baseline risk (95% CrI)
SMSC <sup>18</sup> (real-world study with IPD)	Total	935	191 (20.4)	40.8 (11.2)	631 (67.5)	2.3 (1.4)	20.1 (2.8, 37.5)
AFFIRM <sup>19</sup> (RCT with IPD)	Total	939	359 (38.2)	36.0 (8.3)	657 (70.0)	2.3 (1.2)	36.5 (18.8, 54.1)
	Natalizumab	627	183 (29.2)	35.6 (8.5)	449 (71.6)	2.3 (1.16)	36.9 (19.5, 54.3)
	Placebo	312	176 (56.4)	36.7 (7.8)	208 (66.7)	2.3 (1.19)	35.6 (17.6, 53.7)
CONFIRM <sup>20</sup> (RCT with IPD)	Total	1417	451 (31.8)	37.3 (9.3)	993 (70.1)	2.6 (1.2)	37.2 (18.6, 55.7)
	Dimethyl fumarate	703	185 (26.3)	37.8 (9.4)	495 (70.4)	2.5 (1.2)	36.8 (18.2, 55.3)
	Glatiramer acetate	351	117 (33.3)	36.7 (9.1)	247 (70.3)	2.6 (1.2)	37.4 (17.6, 57.3)
	Placebo	363	149 (41.0)	36.9 (9.2)	251 (69.1)	2.6 (1.2)	37.7 (20.5, 54.9)
DEFINE <sup>21</sup> (RCT with IPD)	Total	1234	394 (31.9)	38.5 (9.0)	908 (73.6)	2.4 (1.2)	36.9 (17.7, 56.0)
	Dimethyl fumarate	826	212 (25.7)	38.5 (9.0)	602 (72.9)	2.4 (1.2)	36.2 (17.2, 55.1)
	Placebo	408	182 (44.6)	38.5 (9.1)	306 (75)	2.5 (1.2)	38.2 (19.0, 57.5)
Bornstein <sup>23</sup> (RCT with AD)	Total	50	30 (60.0)	30.5 (NA)	29 (58.0)	3.1 (NA)	35.6 (19.9, 51.3)
	Glatiramer acetate	25	11 (44.0)	30.0 (NA)	14 (0.6)	2.9 (NA)	NA
	Placebo	25	19 (76.0)	31.1 (NA)	15 (0.6)	3.2 (NA)	NA
Johnson <sup>24</sup> (RCT with AD)	Total	251	186 (74.1)	34.5 (6.4)	184 (73.3)	2.6 (1.3)	30.8 (3.4, 58.1)
	Glatiramer acetate	125	89 (71.2)	34.6 (6.0)	88 (70.4)	2.8 (1.2)	NA
	Placebo	126	97 (77.0)	34.3 (6.5)	96 (76.2)	2.4 (1.3)	NA

Abbreviations: AD, aggregate data; CrI, credible interval; EDSS, Expanded Disability Status Scale; IPD, individual participant data; NA, not available; RCT, randomized clinical trial; sd, standard deviation; SMSC, Swiss Multiple Sclerosis Cohort.

### 3.1 | Notation

Consider a set of treatments  $\mathcal{H}$  each denoted by  $h = 1, 2, \dots, T$ . Let  $y_{ijh}$  denote the dichotomous outcome for individual  $i=1, 2, \dots, n$  under treatment  $h$  in the  $j$ -th trial, and a total of  $ns$  trials. An individual can experience the outcome ( $y_{ijh} = 1$ ) or not ( $y_{ijh} = 0$ ).  $PF_{ijk}$  is the  $k$ -th prognostic factor,  $k = 1, 2, \dots, np$ . The  $np$  prognostic factors are

used to estimate the baseline risk  $R_i$  (independent of treatment), for each participant. The probability of the outcome to occur for individual  $i$  in study  $j$  under treatment  $h$  is denoted by  $p_{ijh}$  and depends on treatment, baseline risk  $R_i$  and the interaction between the baseline risk and the treatment. We use asterisk (\*), to differentiate between the estimations before and after recalibration:  $R_i^*$  indicates the baseline risk before the

re-calibration, estimated using the SMSC, while  $R_i$  indicates the baseline risk after re-calibration, for the RCTs population.

### 3.2 | Stage 1: Development and internal validation of the baseline risk prognostic model

There is plenty of guidance about how to develop and validate a prognostic model.<sup>22,34–36</sup> Good practice involves the use of appropriate model selection methods (or pre-specifying the model), shrinkage in the coefficients to avoid extreme predictions, accounting for missing data and correcting for optimism when the model performance is evaluated internally.

In our approach, we developed the prognostic model using a non-randomized study for the baseline risk,  $R_i^*$ , for each individual  $i$ . We used a logistic mixed-effects model, which accounts for information about the same patient from different cycles. ( $c$ , where  $c = 1, 2, \dots, nc$ ), in a Bayesian framework as

$$\text{logit}(R_i^*) = \beta_0^* + u_{0i}^* + \sum_{k=1}^{np} (\beta_k^* + u_{ki}^*) \times \text{PF}_{ik}. \quad (1)$$

$\beta_0^*$  and  $\beta_k^*$  are the fixed effect intercept and fixed effect slopes respectively, and  $u_{0i}^*$  and  $u_{ki}^*$  are the individual-level random effects intercept and individual-level random effects slopes, which account for information about the same patient from different cycles. A detailed description of the model development and internal validation is available elsewhere.<sup>37</sup>

### 3.3 | Stage 2: Re-calibration of the baseline risk prognostic model for populations included in RCTs

While observational data in stage 1 might lead to better estimation of the prognostic effect of baseline covariates under real-world conditions,<sup>38–41</sup> the predictions for different populations, like this of RCTs, might be less accurate. To estimate parameters that relate to treatment effects and their modification, it is best to use RCT data. The model is described in Stage 3 and uses the baseline risk as a covariate; this baseline risk is best to be as accurate as possible for the RCT population. In this stage 2, we will re-calibrate the baseline risk model  $R_i$  of stage 1.<sup>22,42</sup> We will start with the model as of Equation (1) as estimated using the SMSC data and then use the RCT data to (1) recalibrate the model intercept,

(2) recalibrate the intercept and the overall calibration slope, and (3) recalibrate the intercept, the overall slope, and re-estimate some of the regression coefficients.<sup>22,42</sup>

Recalibrating only the intercept ensures that the average predicted baseline risk is equal to average observed baseline risk in RCTs.<sup>42</sup> The recalibrated baseline risk  $R_i$  can be estimated by plugging-in the estimated slopes  $\beta_k^*$  from stage 1 (Equation 1) and then re-estimate the intercept  $\beta_0$ , by fitting

$$\text{logit}(R_i) = \beta_0 + \text{logit}(R_i^*) \quad (2)$$

to the RCTs data. The intercept  $\beta_0$  could be assumed exchangeable ( $\beta_0 \sim N(b_0, \sigma_{b_0}^2)$ ), or common ( $\beta_0 = b_0$ ) across studies.

Another option is to recalibrate the intercept and the overall calibration slope,  $\beta_{\text{overall}}$ .<sup>42</sup>

This will also update the overall effect of the prognostic factors for the RCTs setting. We first estimate the uncalibrated predictions  $R_i^*$  for the RCT population, and then we estimate the following model

$$\text{logit}(R_i) = \beta_0 + \beta_{\text{overall}j} \times \text{logit}(R_i^*), \quad (3)$$

where the intercept and the overall regression coefficient of  $\text{logit}(R_i^*)$  could be assumed exchangeable ( $\beta_0 \sim N(b_0, \sigma_{b_0}^2), \beta_{\text{overall}j} \sim N(b_{\text{overall}}, \sigma_{b_{\text{overall}}}^2)$ ), or common ( $\beta_0 = b_0, \beta_{\text{overall}j} = b_{\text{overall}}$ ) across studies. The recalibrated predicted risk score  $R_i$  is obtained from Equation (3) after estimating  $b_0$  and  $b_{\text{overall}}$  via the RCTs with IPD.

A more comprehensive option is to re-calibrate the intercept, the overall slope (as above) and in addition re-estimate some of the regression coefficients as needed.<sup>42</sup> The re-estimated baseline risk for RCT patients,  $R_i$ , will be finally estimated as:

$$\text{logit}(R_i) = \beta_0 + \sum_{k=1}^{np} \beta_{kj} \times \text{PF}_{ik}, \quad (4)$$

where  $\beta_0$  and  $\beta_{kj}$  are the recalibrated intercept and regression coefficients, and as before can be assumed exchangeable ( $\beta_0 \sim N(b_0, \sigma_{b_0}^2), \beta_{kj} \sim N(b_k, \sigma_{b_k}^2)$ ), or common ( $\beta_0 = b_0, \beta_{kj} = b_k$ ) across studies.<sup>42</sup>

The common-effects assumption for  $\beta_0, \beta_{\text{overall}j}$ , and  $\beta_{kj}$ , ignores trial differences, and could be used only in cases

where RCTs were designed using the same or similar protocols and inclusion criteria. A viable option that might be relevant in most cases is to assume random effects across studies. The parameters could be also estimated independently (e.g., each  $\beta_{0j}, \beta_{\text{overall}j},$  and  $\beta_{kj}$  are given a prior distributions); however, this would lead to different baseline risks per study, which would pose challenges in the baseline risk estimation for a new patient with a variables' profile not included in one of the available studies.

It is important to note that the development of the baseline risk and its recalibration (steps 1 and 2) do not aim to predict the outcome risk accurately. Instead, they aim to reduce dimensionality by synthesizing baseline information into a single variable. We followed the recommendation outlined in the predictive approaches to treatment effect heterogeneity (PATH) statement.<sup>11</sup> and developed the baseline risk model (steps 1 and 2) using the entire trial population blinded to treatment assignments. In this context, the variable  $R_i$  represents the baseline risk while being unaware of the treatment allocation.<sup>3,11,17,43</sup>

In the application, we use the re-calibration method associated with the best model's calibration (i.e., the agreement between the observed outcome's proportions and the predicted probabilities) and discrimination ability (i.e., area under the curve [AUC]).

### 3.4 | Stage 3: Network meta-regression with individual and aggregate data using the baseline risk as prognostic factor and effect modifier

In the third stage, we used the calibrated  $\text{logit}(R_i)$  from stage 2 as covariate in a network meta-regression model.<sup>23</sup> We extended the meta-regression model suggested by Saramago et al. to combine IPD and AD in a network of trials comparing multiple treatments.<sup>23</sup> In the first part, we modeled studies with IPD

$$y_{ijt} \sim \text{Bernoulli}(p_{ijh}),$$

$$\text{logit}(p_{ijh}) = \begin{cases} u_j + g_{0j} \times (\text{logit}(R_i)) & \text{if } h = h_{\text{ref},j} \\ u_j + d_{jh_{\text{ref},j}h} + (g_{0j} + g_{jh_{\text{ref},j}h}^W) \times (\text{logit}(R_i)) + (g_{h_{\text{ref},j}h}^B - g_{jh_{\text{ref},j}h}^W) \times \overline{\text{logit}(R)}^j, & \text{if } h \neq h_{\text{ref},j} \end{cases}, \quad (5)$$

where  $\overline{\text{logit}(R)}^j$  is the mean logit baseline risk from all patients in study  $j$ , and each study  $j$  has a reference treatment  $h_{\text{ref},j} \in \mathcal{H}$ .

The parameters of interest are the relative treatment effects  $d_{jh_{\text{ref},j}h}$ . We estimated independently the nuisance parameters  $u_j$  for each study (i.e., the log odds of experiencing the outcome under the study's reference treatment). The coefficients  $g_{0j}$  measure the prognostic impact of baseline risk and can be assumed independent, exchangeable ( $g_{0j} \sim N(\gamma_0, \sigma_{\gamma_0}^2)$ ), or common ( $g_{0j} = \gamma_0$ ) across studies. The regression coefficients  $g_{jh_{\text{ref},j}h}^W$  measure how the baseline risk of a patient modifies the treatment effect within each study; they can be combined across studies assuming random ( $g_{jh_{\text{ref},j}h}^W \sim N(G_{h_{\text{ref},j}h}^W, \sigma_{G^W}^2)$ ) or common ( $g_{jh_{\text{ref},j}h}^W = G_{h_{\text{ref},j}h}^W$ ) effects, where  $G_{h_{\text{ref},j}h}^W = \gamma_h^W - \gamma_{h_{\text{ref},j}}^W$  with  $\gamma_{\text{ref}}^W = 0$  for an overall reference treatment. Similarly, the between-studies effect modification parameters  $g_{h_{\text{ref},j}h}^B$  measure how the mean baseline risk of each study modifies the relative treatment effect.

In the second part, we synthesize information from studies that report only AD. The likelihood of the observed data in AD studies is

$$r_{jt} \sim \text{Binomial}(p_{jh}, n_{jh}),$$

where  $r_{jh}, n_{jh}, p_{jh}$  denote the number of patients experiencing the outcome of interest, the total number of randomized individuals and the probability of experiencing the outcome, in study  $j$  in treatment arm  $h$ , respectively.

Then, we model the relative treatments effects using the average study-specific baseline risk  $\overline{\text{logit}(R)}^j$

$$\text{logit}(p_{jh}) = \begin{cases} u_j, & \text{if } h = h_{\text{ref},j} \\ u_j + d_{jh_{\text{ref},j}h} + g_{h_{\text{ref},j}h}^B \times \overline{\text{logit}(R)}^j, & \text{if } h \neq h_{\text{ref},j} \end{cases}. \quad (6)$$

To estimate the average baseline risk  $\overline{\text{logit}}(R)^j$  we simulated pseudo-IPD using a multivariate normal distribution with means equal to the reported mean covariate values ( $\overline{\text{PF}}_{kj}$ ), and variance–covariance matrix calculated using the reported standard deviations and correlations between covariates estimated from the RCTs with IPDs. In this way, we might attenuate potential ecological bias in the estimation of  $\overline{\text{logit}}(R)^j$ .

The mean values of some of the prognostic factors might not be reported in the original studies. In that case, we used imputations to allow studies with partial information on covariates to be included in the meta-regression model, as previously described by Hemming et al. (described in [Appendix](#), Supporting information).<sup>44</sup>

In the third part, the relative treatment effects,  $d_{jh_{\text{ref},j}h}$ , can be combined across studies in a random-effect ( $d_{jh_{\text{ref},j}h} \sim N(D_{jh_{\text{ref},j}h}, \sigma_D^2)$ ) or common-effect ( $d_{jh_{\text{ref},j}h} = D_{jh_{\text{ref},j}h}$ ) across studies assuming consistency  $D_{jh_{\text{ref},j}h} = \delta_h - \delta_{h_{\text{ref},j}}$  where  $\delta_{\text{ref}} = 0$ . Finally, the consistency equations for the within and between studies effect modification parameters  $g_{jh_{\text{ref},j}h}^B$  are  $G_{jh_{\text{ref},j}h}^B = \gamma_h^B - \gamma_{h_{\text{ref},j}}^B$  and  $\gamma_{\text{ref}}^B = 0$ .

The difference between  $g_{jh_{\text{ref},j}h}^W$  and  $g_{jh_{\text{ref},j}h}^B$  represents an estimate of ecological bias (i.e., the difference between across-study associations and within-study associations, due to study-level confounding).<sup>45</sup> To ensure we do not introduce bias, the within-study effect modification ( $g_{jh_{\text{ref},j}h}^W$ ) is estimated through the IPD studies alone (Equation 5), whereas both IPD and AD studies are used for the between-study effect modification estimation ( $g_{jh_{\text{ref},j}h}^B$ ).<sup>23,45</sup>

The aim of Stage 3 is to estimate accurately the treatment effects ( $\delta_h$ ), adjusted for the baseline risk ( $R_i$ ), and the interactions between the treatments and the baseline risk within ( $\gamma_h^W$ ) and between ( $\gamma_h^B$ ) studies.

### 3.5 | Making treatment-specific outcome predictions

To predict the probability  $p_{i_{\text{new}}h}$  of the outcome in a new patient  $i_{\text{new}}$  under treatment  $h$ , we follow a series of steps.

#### 3.5.1 | Step 1—Estimation of baseline risk for a new patient

We estimate the  $\text{logit}(R_{i_{\text{new}}})$ , which represents the log-odds of the baseline risk for the new patient, using the

coefficients  $b_0, b_k$  obtained from stage 2, if we aim to make predictions for RCTs populations or using the coefficients  $\beta_0^*, \beta_k^*$  obtained from stage 1, if we aim to make predictions for real-world population.

#### 3.5.2 | Step 2—Estimation of reference treatment parameters

We need to estimate from a population similar to those that we want to make predictions:

- the logit-probability of the outcome under the reference treatment (placebo, in our example)—denoted as  $a$ .
- the regression coefficient of  $\text{logit}(R_{i_{\text{new}}})$  under the reference treatment—denoted as  $\gamma$ , and can be interpreted as the average change in the logit-probability of relapse within 2 years for a one unit increase in the logit-transformed baseline risk, specifically for patients receiving the reference treatment.

If our aim is to make predictions for RCT populations, then placebo arms from RCTs can be used to estimate  $a$ , and  $\gamma$  using a meta-regression model, which (under common treatment effects assumption across studies) would be:

$$\text{logit}(p_{I,\text{placebo}}) = a + \gamma \times \text{logit}(R_I). \quad (7)$$

If we aim to make predictions in a real-world population, then untreated patients from a registry or a cohort study can be used to estimate  $a$ , and  $\gamma$  as in Equation (7).

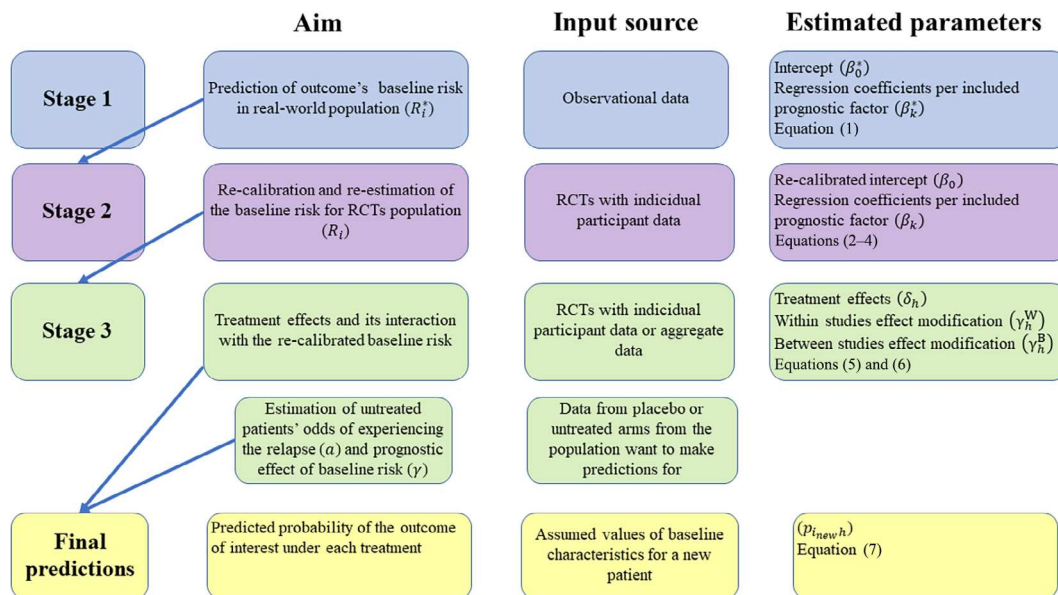
#### 3.5.3 | Step 3—Estimation of the mean logit baseline risk across all individuals

If our aim is to make predictions for RCT populations, then all patients from several RCTs can be used to estimate the mean of logit baseline risk across all individuals (i.e., mean of  $\text{logit}(R_i)$ ), denoted as  $\overline{\text{logit}}(R)$ . If we aim to make predictions in a real-world population, then all patients from a registry or a cohort study can be used for the  $\overline{\text{logit}}(R)$  estimation.

#### 3.5.4 | Step 4—Final prediction

Assuming common treatment effects, we use the following equation





**FIGURE 1** Schematic presentation of each stage's aim, suggested data design and type, and estimated parameters. RCTs, randomized clinical trials.

$$\text{logit}(p_{i_{new,h}}) = a + \delta_h + (\gamma + \gamma_h^W) \times \text{logit}(R_{i_{new}}) + (\gamma_h^B - \gamma_h^W) \times \left( \overline{\text{logit}(R)} \right) \tag{8}$$

The values for  $\delta_h, \gamma_h^W$ , and  $\gamma_h^B$ , are those estimated in the third stage of the network meta-regression prognostic model (Equation 5).  $a$  and  $\gamma$  are the reference treatment parameters as estimated in Equation (7), and  $\left( \overline{\text{logit}(R)} \right)$  is the mean logit baseline risk estimated from a population similar to the one we aim to make predictions. Under the random effects assumption,  $\text{logit}(p_{i_{new,h}})$  would be normally distributed with mean as in Equation (8), and a variance-covariance matrix determined by the variance-covariance matrix of  $\delta_h$ , and  $\gamma_h^W$ .

Figure 1 presents a schematic presentation of the aim, data, and parameters of each stage of the approach. Information about the studies used in the example is also presented.

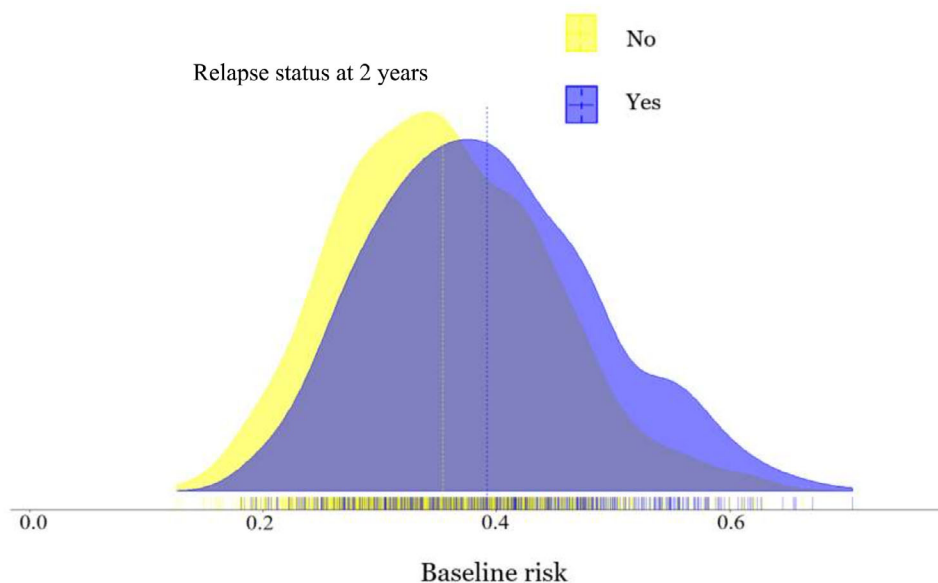
#### 4 | APPLICATION: HETEROGENEOUS EFFECTS OF TREATMENTS FOR RRMS

We developed the prognostic model using the SMSC data.<sup>28</sup> The development of the prognostic model in stage one has been previously published and is implemented as Shiny app in <https://cinema.ispm.unibe.ch/shinies/rrms/>.<sup>37</sup>

We first selected the prognostic factors via a review of the literature.<sup>46–51</sup> We included all prognostic factors that were included in at least two previously published prognostic models. The model includes eight prognostic factors: age, sex, EDSS, prior or current treatment (yes or no), months since last relapse, disease duration, number of relapses in the previous 2 years, number of gadolinium enhanced lesions. We then fitted a logistic mixed-effects regression model in a Bayesian framework accounting for correlations induced by individuals contributing data to more than one cycle. The SMSC includes 1752 observations from 2-years cycles of 935 patients, and 302 of those patients experienced at least one relapse. The full model had 22 degrees of freedom (for 10 predictors with random intercept and slope) and the number of events per variable was 13.7. To shrink the coefficients of the regression and avoid extreme predictions, we used Laplace prior distributions.<sup>52</sup> We used multiple imputation to account for missing covariate data.<sup>53,54</sup> After internal validation, the bootstrap optimism-corrected AUC was 0.65 and the bootstrap optimism-corrected calibration slope 0.91. The calibration plot and the evaluation of the model's clinical usefulness are presented elsewhere.<sup>37</sup> The model's accuracy and clinical performance are overall suggesting a useful prediction model.

We then re-calibrated the prognostic model for the RCT setting (stage 2). All three RCTs used for the re-calibration of the baseline risk model in stage 2 had similar protocols developed by the same company, and had participants with similar baseline characteristics as

**FIGURE 2** The distribution of predicted baseline risk of any relapse within the next 2 years for individuals by relapse status in the randomized clinical trials dataset (stage 2). The dashed lines indicate the mean of predicted baseline risk for individuals who did experience a relapse (purple) and for those who did not (yellow).



presented in Table 1. Therefore, we assumed common-effects across the three RCTs for estimating the intercept (in Equations 2–4), the overall regression coefficient (in Equation 3), as well as the regression coefficients of each prognostic factor (Equation 4). Alternatively, when studies share different designs, protocols, and inclusion criteria, random effects across studies should be assumed in stage 2. We assessed the predictions of the developed re-calibrated models in a calibration plot with loess smoother (Appendix Figure 1, Supporting information). The re-calibration method resulting in the highest optimism-corrected AUC (AUC = 0.61) and the best optimism-corrected calibration ( $c$ -slope = 1.002 and  $c$ -intercept = 0.004) was “the re-calibration and selective re-estimation” approach (Equation 4); the other two methods resulted in optimism-corrected AUC = 0.58. The “re-calibration of intercept” method resulted to an optimism-corrected  $c$ -slope = 0.85 and  $c$ -intercept = −0.08, and the “re-calibration of intercept and overall slope” method resulted to an optimism-corrected  $c$ -slope = 0.994 and  $c$ -intercept = 0.005. Based on the existing literature, risk models with a low predictive ability (0.6–0.65) are often adequate to detect risk-based heterogeneous treatment effects.<sup>4,18</sup> Therefore, we selected the baseline risk from “the re-calibration and selective re-estimation” method to use in the next stage of the risk modeling approach. Appendix Table 1, Supporting information presents the re-calibrated regression coefficients for each prognostic factor.

In Figure 2, we show the distributions of the predicted baseline risk by relapsing status in the populations included in the three RCTs. The overall mean predicted baseline risk was 36.8% (95% credible interval [CrI] 36.4%–37.2%). The overlap in the distributions was large,

as reflected by the low AUC. For patients who experienced a relapse, the mean predicted risk was 39.2% (95% CrI 38.5%–39.8%) whereas for patients who did not, it was 35.4% (95% CrI 34.9%–35.9%).

The predicted baseline risk in the RCT populations was then used in the network-meta-regression model (Equations 5 and 6, Stage 3). Because only two AD studies were available, we assumed that  $g_{jh_{ref,j}h}^W = g_{jh_{ref,j}h}^B$  to enable model convergence. We also assumed that study-specific relative treatment effects do not have any residual heterogeneity beyond what is already captured by differences in baseline risk. As the heterogeneity variance was not well estimated with five studies, we assumed common relative treatment effects ( $d_{jh_{ref,j}h} = D_{h_{ref,j}h}$ ) and common-effect modification across studies ( $g_{jh_{ref,j}h}^W = G_{h_{ref,j}h}^W$ ). We also assumed common coefficients for the prognostic effect of the baseline risk ( $g_{0j} = \gamma_0$ ), as all three studies with IPD data were very similar in terms of design and patient characteristics.

None of the two AD studies provided information about the number of patients with prior treatment, gadolinium enhanced lesions, and months since last relapse; we performed imputations as described in the Appendix A, Supporting information. Then, we created two pseudo-IPDs (one for each AD study) and estimated the baseline risk for each patient in the pseudo-IPD datasets (presented in Appendix A, Supporting information). The estimated mean baseline risk for each study is presented in Table 1.

Table 2 shows the estimated parameters from the network meta-regression model (Stage 3). In addition, we performed a sensitivity analysis excluding the AD studies.

**TABLE 2** Estimated parameters from network meta-regression model including the logit-risk as covariate (Stage 3).

Estimated parameters from network meta-regression model	Mean (95% CrI)
OR of relapsing for one unit increase in logit-risk ( $e^{\gamma_0}$ )	2.72 (2.02, 3.70)
OR of relapsing under DF versus placebo ( $e^{\delta_{DF}}$ )	0.39 (0.25, 0.59)
OR of relapsing under GA versus placebo ( $e^{\delta_{GA}}$ )	0.41 (0.22, 0.77)
OR of relapsing under N versus placebo ( $e^{\delta_N}$ )	0.21 (0.12, 0.34)
OR of relapsing under DF versus placebo for one unit increase in logit-risk ( $e^{\gamma_{DF}^W}$ )	0.84 (0.44, 1.62)
OR of relapsing under GA versus placebo for one unit increase in logit-risk ( $e^{\gamma_{GA}^W}$ )	0.64 (0.27, 1.53)
OR of relapsing under N versus placebo for one unit increase in logit-risk ( $e^{\gamma_N^W}$ )	0.60 (0.30, 1.21)

Abbreviations: CrI, credible interval; DF, dimethyl fumarate; GA, glatiramer acetate; N, natalizumab, OR, odds ratio.

Appendix Table 2, Supporting information shows the results of the network meta-regression model (Stage 3), when including and excluding the AD studies; the results are similar in both cases. The estimated values of  $\gamma_0$  indicate that baseline risk is an important prognostic factor for relapse. We first make predictions for the RCT populations and hence we estimate  $a$ , and  $\gamma$  by synthesizing data from placebo individuals across the three RCTs with IPD and  $\overline{\text{logit}(R)}$  as the mean of  $\text{logit}(R_i)$  across all individuals in RCTs with IPD. The treatment effects as a function of the baseline risk are shown in Figure 3. Appendix Figure 2, Supporting information presents the final predictions with their 95% CrIs. Natalizumab gives the lowest probability of relapsing over almost the entire baseline risk range. However, its advantage over dimethyl fumarate for patients with low baseline risk (below 30%, on average) is very small. We also assessed the apparent prediction model's accuracy using the calibration plot with loess smoother (Appendix Figure 3, Supporting information). The results indicate that the model accurately predicts the probability of relapse within the next 2 years. While optimism-corrected internal validation is considered optimal for assessing model performance, it is essential to clarify that the primary focus of this work is to illustrate and demonstrate the methodology we have presented.

Additionally, we performed a sensitivity analysis comparing the final predictions from Stage 3 under all three recalibration methods (Appendix Table 3 and Figure 4, Supporting information). All three recalibration methods lead to similar final predictions. On average, natalizumab

minimizes the predicted probability of relapsing within the next 2 years (about 7%–12% mean absolute difference compared to dimethyl fumarate). For low-risk patients (baseline risk  $\leq 30\%$ ), the probability to relapse is similar under dimethyl fumarate and natalizumab. For high-risk patients (baseline risk  $\geq 50\%$ ), natalizumab is the drug that minimizes the predicted probability of relapsing within the next 2 years (about 13%–19% mean absolute difference compared to dimethyl fumarate).

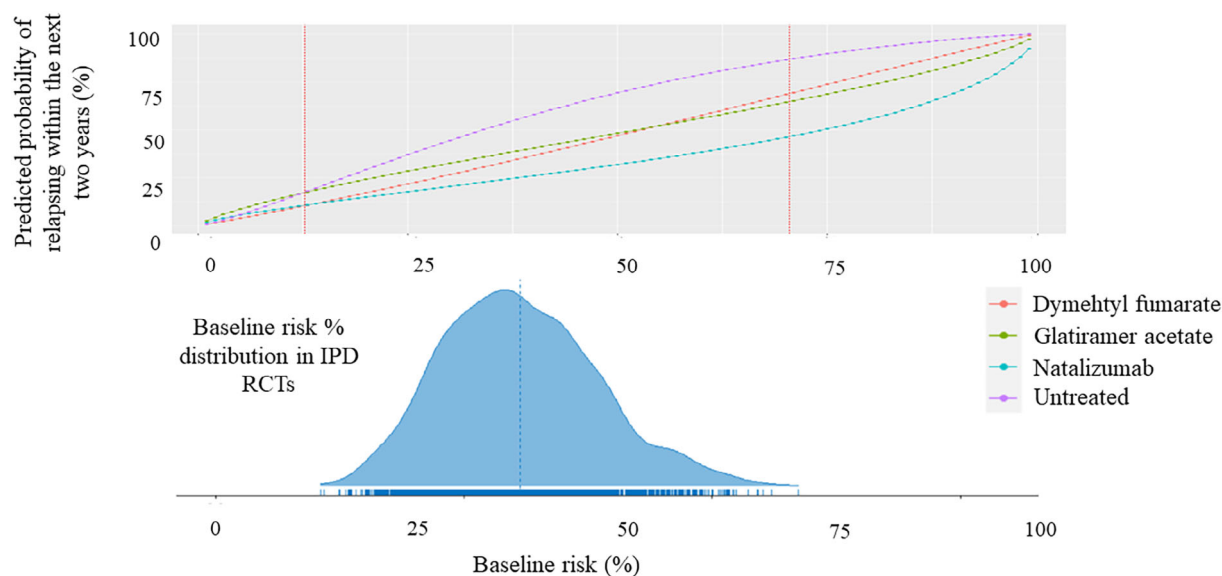
To make predictions for the Swiss real-world population, we estimate  $a$  as the logit-probability of relapse in untreated patients in the SMSC and  $\overline{\text{logit}(R)}$  as the mean of  $\text{logit}(R_i)$  across all individuals in the SMSC;  $\gamma$  was also estimated using SMSC. The results for the SMSC population are presented in Appendix Figure 5, Supporting information. Often the patients' baseline disease condition is more severe in RCTs than in observational studies, and hence as expected, the distribution of baseline risk is different between the SMSC and the RCTs. However, the relative ranking of therapies for a given baseline risk does not deviate from this of RCTs (presented in Figure 3). As in the RCTs population, the advantage of natalizumab over dimethyl fumarate for patients with low baseline risk is non-existing (Appendix Figure 5, Supporting information). An interactive version of Figure 3 and Appendix Figure 5, Supporting information has been implemented in <https://cinema.ispm.unibe.ch/shinies/srrms/>.

Table 3 summarizes the information in Figure 3 for patients at baseline risk below 30% (low risk) or more than 50% (high risk). These cut-offs were chosen arbitrarily for illustrative purposes. For high-risk patients (8.5% of patients in RCTs), the risk difference for relapse between natalizumab and dimethyl fumarate is 19% favoring natalizumab. For low-risk patients (25% of patients in RCTs), the risk difference between natalizumab and dimethyl fumarate is 1.4%.

## 5 | DISCUSSION

We developed a three-stage network meta-analysis approach, where data from different sources and study designs can be synthesized to make predictions for heterogeneous treatment effects. We exemplified our method by predicting the probability of relapse under three active treatments and placebo in patients with RRMS, we made the code available ([https://github.com/esm-ispm-unibe-ch/ThreeStageModel\\_RRMS](https://github.com/esm-ispm-unibe-ch/ThreeStageModel_RRMS)), and we created an online tool to show the predictions in an interactive way (<https://cinema.ispm.unibe.ch/shinies/srrms/>).

Central to our work is the risk modeling approach. The main advantage of the risk modeling method is that it reduces dimensionality by summarizing all relevant



**FIGURE 3** Predicted probability of relapsing within the next 2 years as a function of the baseline risk (Stage 3) into the randomized clinical trials (RCTs) population. The *x*-axis shows the baseline risk of relapsing within the next 2 years (after re-calibration, stage 2) and the *y*-axis shows the predicted probability to relapse within the next 2 years under each one of the available treatments. Between the two red vertical dashed lines are the baseline risk values observed in the three RCTs with individual participant data.<sup>29–31</sup> The distribution of the baseline risk in these three trials is presented at the bottom of the graph. IPD, individual participant data.

**TABLE 3** Predicted % average risk difference and odds ratio of each active treatment versus placebo in the populations in randomized clinical trials. Results are shown for all patients, and for two baseline risk groups.

Treatment effects	Treatment	All patients	Baseline risk <30% low-risk patients	Baseline risk >50% high-risk patients
Risk difference of drug versus placebo (95% CrI)	Dimethyl fumarate	38.2 (26.1, 50.4)	17.6 (10.5, 29.8)	58.1 (37.9, 73.9)
	Glatiramer acetate	41.1 (25.3, 57.6)	24.5 (13.8, 42.3)	56.6 (31.5, 77.0)
	Natalizumab	27.2 (16.1, 40.6)	15.2 (8.6, 26.2)	39.0 (20.2, 59.4)
Odds ratio of drug versus placebo (95% CrI)	Dimethyl fumarate	0.43 (0.32, 0.57)	0.50 (0.35, 0.71)	0.36 (0.25, 0.51)
	Glatiramer acetate	0.53 (0.35, 0.83)	0.78 (0.49, 1.30)	0.34 (0.20, 0.61)
	Natalizumab	0.28 (0.20, 0.39)	0.43 (0.29, 0.62)	0.17 (0.10, 0.26)

Abbreviation: CrI, credible interval.

baseline variables in one “baseline risk” variable. The risk of overfitting in the final prediction model is then low and model selection methods and shrinkage are needed when developing the baseline risk model, but not when making the final predictions. Risk modeling approach has been originally introduced as a method to analyze a single randomized trial,<sup>3,8,16–20</sup> then extended to meta-analysis<sup>3,8,55</sup> and more recently to network meta-analysis<sup>21</sup> of randomized trials. However, none of these approaches examined combining and making the best use of all available data sources. Observational studies reflect better the real-world populations and

conditions<sup>38–41</sup> and this is why we used a cohort in the first stage of our approach to develop a model that predicts the baseline risk. In addition, we combined AD and IPD to increase the power and precision of the estimated treatment effects.

Our methodology relies on two main sets of assumptions: those of network meta-analysis<sup>2</sup> and risk modeling assumptions.<sup>11</sup> Within the network meta-analysis framework, we emphasize the transitivity assumption, which suggests that patient populations, study designs, and other factors that can modify the treatment effect should closely align across groups of studies that compare

different interventions. Examination of the standard assumptions underlying meta-analysis and meta-regression are also required: normality in the distribution of random effects, knowledge of the variance of the treatment effects, etc.<sup>56</sup> In addition, the risk modeling approach assumes that the variables composing the risk score comprehensively capture both prognosis and effect modification.<sup>11,57</sup> Evaluating this assumption can be challenging, particularly when the outcome is insufficiently studied or when there is a scarcity of prognostic studies on the subject. These assumptions are essential foundations of our approach, ensuring the validity and reliability of the results. The approach has several caveats. It requires that at least one IPD dataset per intervention is available. The access to IPD data entails many challenges and difficulties described in detail elsewhere.<sup>58–60</sup> We implemented the three-stage approach by fitting three separate models instead of developing a single Bayesian model. We used an existing prognostic model for the baseline risk (stage 1), and we re-calibrated it (stage 2) within a frequentist setting to take advantage of the software's re-calibration options. Consequently, uncertainty was not accounted between the different stages and the results from stage three might be over-precise. In addition, the imputation method used, although it allows the use of AD studies even if study-level covariates are missing, may not be the optimal one. Other methods, like advanced multiple imputations techniques for study-level characteristics, may be used.<sup>61</sup> Finally, in the RRMS application, we used common treatment effects model (Stage 3) to enable model convergence, because of the small number of studies. This assumption can be relaxed if more studies are available.

The implementation of our approach in the RRMS example shows that several patient characteristics influence the baseline risk of relapse, which in turn modifies the effect of treatments. Natalizumab appears to be the optimal treatment (i.e., minimizes the predicted probability of relapsing) over almost the entire baseline risk range. However, its advantage over dimethyl fumarate for patients with low baseline risk is very small or non-existing. Dimethyl fumarate might be the optimal treatment for these patients, as natalizumab is a drug considered less safe.<sup>26,27</sup> The results are in agreement with those of a recent published work,<sup>21</sup> as well as aligned with expert opinions in neurology. The optimism-corrected discrimination of the prediction model for the baseline risk (stage 2) was small (AUC = 0.61) but sufficient for our aim. The discriminative ability of the existing prognostic models for relapses in MS is generally low (less than 65%), indicating that relapses might be associated with unknown factors.<sup>21,37,49,62</sup> Second, it has been shown that models with

low AUC can still be useful when their predictions are used as potential effect modifiers of treatment.<sup>4,18,63</sup> This was the case in our application where we showed clinically meaningful differences between the interventions for different levels of the baseline risk (Figure 3). Note that the findings of this model, are as expected by practitioners and as described in international guidelines, namely that natalizumab is to be administered as a second line treatment. Prognostic scores from models with low AUC were previously used as effect modifiers; see for example, the Thrombolysis in Myocardial Infarction and the CHADS2 risk score; both were powerful in detecting the heterogeneous treatment effects of via a risk modeling approach.<sup>64–68</sup>

The application in the example of RRMS is used as an example to illustrate the methodology, showing the potential of our approach, which can make prediction of individualized treatment effects in RCTs and real-world populations; however, it is not ready for use in clinical practice. Decision-making tools need bootstrap (or cross-validation) internal and external validation and need to show evidence about all relevant treatment options, before they are considered for use.<sup>69</sup> In addition, we focused primarily on validating the model's predicted probability to relapse; future research could explore more sophisticated validation techniques for validating the treatment benefit.<sup>70–73</sup> Finally, similar to the well-known phases of clinical research in drug development, Heinze et al. define four phases of methodological research.<sup>74</sup> The present article can be considered as a phase I development, where an idea is introduced and, based on the fact that its components are well established in the evidence synthesis and prognostic research field, it is expected to provide valid results. However, phase II and III studies with extensive simulation scenarios are needed before the proposed models are employed in practice.

The presented approach offers many opportunities for further development. Several bias-adjusted methods have been proposed to combine non-randomized data and RCTs to estimate average treatment effects in a meta-analysis framework.<sup>75,76</sup> Some of them use the baseline risk to adjust for selection bias in the real-world data in a meta-analysis framework.<sup>75</sup> The approach we presented could be further extended by using observational data to inform not only the baseline risk in stage 1, but also the relative treatment effects (in Stage 3) using appropriate bias-adjusted modeling.<sup>75,76</sup> Evidence about treatment cost and safety could be incorporated to extend the model further and to better inform clinical decision-making. Besides, it is possible that other study-level characteristics, like the year of randomization and the risk of bias, may also influence the treatment effects. Such variables can be added to the network meta-regression model, if

the number of studies permits. Finally, the implementation of the three stages into a single Bayesian model will allow naturally to incorporate uncertainty from all stages in the final result and avoid spuriously overprecise conclusions.

## 6 | CONCLUSION

The proposed approach combines all relevant evidence sources and can be applied to estimate individualized predictions of treatment effects for any health condition. Consequently, it has the potential to assist clinical practice and decision-making toward treatment recommendations and precision medicine.

### AUTHOR CONTRIBUTIONS

**Konstantina Chalkou:** Conceptualization; data curation; formal analysis; methodology; resources; software; supervision; visualization; writing – original draft; writing – review and editing. **Tasnim Hamza:** Formal analysis; methodology; writing – review and editing. **Pascal Benkert:** Conceptualization; data curation; writing – review and editing. **Jens Kuhle:** Conceptualization; data curation; writing – review and editing. **Chiara Zecca:** Conceptualization; data curation; writing – review and editing. **Gabrielle Simoneau:** Conceptualization; data curation; writing – review and editing. **Fabio Pellegrini:** Conceptualization; data curation; writing – review and editing. **Andrea Manca:** Conceptualization; methodology; writing – review and editing. **Matthias Egger:** Conceptualization; methodology; writing – review and editing. **Georgia Salanti:** Conceptualization; methodology; supervision; writing – review and editing.

### ACKNOWLEDGMENTS

Konstantina Chalkou, Tasnim Hamza, Andrea Manca, and Georgia Salanti are funded by the European Union's Horizon 2020 research and innovation program under grant agreement no. 825162. Matthias Egger was supported by the Swiss National Science Foundation (grant 189498). The authors thank Suvitha Subramanian for her assistance on cleaning the data. The authors would like to thank Suvitha Subramanian for her support and help regarding the SMSC data cleaning. The Swiss MS Cohort study received funding from the Swiss MS Society and grant funding from Biogen, Celgene, Merck, Novartis, Roche, and Sanofi. Open access funding provided by Universitat Bern.

### CONFLICT OF INTEREST STATEMENT

Konstantina Chalkou, Tasnim Hamza, Pascal Benkert, Jens Kuhle, Matthias Egger, Andrea Manca, and Georgia

Salanti declare that they have no conflict of interest with respect to this paper, JK received speaker fees, research support, travel support, and/or served on advisory boards by Swiss MS Society, Swiss National Research Foundation (320030\_189140/1), University of Basel, Progressive MS Alliance, Bayer, Biogen, Celgene, Merck, Novartis, Octave Bioscience, Roche, Sanofi. Gabrielle Simoneau and Fabio Pellegrini are employees of and hold stocks/stock options in Biogen. Ente Ospedaliero Cantonale (employer) received compensation for Chiara Zecca's speaking activities, consulting fees, or research grants from Almirall, Biogen Idec, Bristol Meyer Squibb, Genzyme, Lundbeck, Merck, Novartis, Teva Pharma, Roche.

### DATA AVAILABILITY STATEMENT

The data that support the findings of this study were made available from Biogen International GmbH and the Swiss Multiple Sclerosis Cohort (SMSC). Restrictions apply to the availability of these data, which were used under license for this study.

### FUNDING STATEMENT

This research was performed as part of the HTx project. The project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 825162. This dissemination reflects only the authors' view and the commission is not responsible for any use that may be made of the information it contains.

### ORCID

Konstantina Chalkou  <https://orcid.org/0000-0001-9718-021X>

Tasnim Hamza  <https://orcid.org/0000-0002-4700-6990>

Pascal Benkert  <https://orcid.org/0000-0001-6525-8174>

Chiara Zecca  <https://orcid.org/0000-0002-9990-3431>

Andrea Manca  <https://orcid.org/0000-0001-8342-8421>

Matthias Egger  <https://orcid.org/0000-0001-7462-5132>

### REFERENCES

1. Cipriani A, Higgins JPT, Geddes JR, Salanti G. Conceptual and technical challenges in network meta-analysis. *Ann Intern Med.* 2013;159:130-137.
2. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Res Synth Methods.* 2012;3:80-97.
3. Kent DM, Steyerberg E, van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *BMJ.* 2018;363:k4245.
4. Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Med Res Methodol.* 2006;6:18.

5. Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. *BMJ*. 1995;311:1356-1359.
6. Schork NJ. Personalized medicine: time for one-person trials. *Nature*. 2015;520:609-611.
7. Gong X, Hu M, Basu M, Zhao L. Heterogeneous treatment effect analysis based on machine-learning methodology. *CPT Pharmacomet Syst Pharmacol*. 2021;10:1433-1443.
8. Rekkas A, Paulus JK, Raman G, et al. Predictive approaches to heterogeneous treatment effects: a scoping review. *BMC Med Res Methodol*. 2020;20:264.
9. 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.
10. Belias M, Rovers MM, Reitsma JB, Debray TPA, Int'Hout J. Statistical approaches to identify subgroups in meta-analysis of individual participant data: a simulation study. *BMC Med Res Methodol*. 2019;19:183.
11. Kent DM, Paulus JK, van Klaveren D, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement. *Ann Intern Med*. 2020;172:35-45.
12. Tanniou J, van der Tweel I, Teerenstra S, Roes KCB. Estimates of subgroup treatment effects in overall nonsignificant trials: to what extent should we believe in them? *Pharm Stat*. 2017;16:280-295.
13. Van Calster B, van Smeden M, De Cock B, Steyerberg EW. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: simulation study. *Stat Methods Med Res*. 2020;29:3166-3178.
14. Riley RD, Snell KIE, Martin GP, et al. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *J Clin Epidemiol*. 2021;132:88-96.
15. Varadhan R, Segal JB, Boyd CM, Wu AW, Weiss CO. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. *J Clin Epidemiol*. 2013;66:818-825.
16. Kent DM, Nelson J, Dahabreh IJ, Rothwell PM, Altman DG, Hayward RA. Risk and treatment effect heterogeneity: reanalysis of individual participant data from 32 large clinical trials. *Int J Epidemiol*. 2016;45:2075-2088.
17. Burke JF, Hayward RA, Nelson JP, Kent DM. Using internally developed risk models to assess heterogeneity in treatment effects in clinical trials. *Circ Cardiovasc Qual Outcomes*. 2014;7:163-169.
18. Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials*. 2010;11:85.
19. Kozminski MA, Wei JT, Nelson J, Kent DM. Baseline characteristics predict risk of progression and response to combination medical therapy for benign prostatic hyperplasia. *BJU Int*. 2015;115(2):308-318.
20. Sussman JB, Kent DM, Nelson JP, Hayward RA. Improving diabetes prevention with benefit based tailored treatment: risk based reanalysis of Diabetes Prevention Program. *The BMJ*. 2015;350:h454.
21. Chalkou K, Steyerberg E, Egger M, Manca A, Pellegrini F, Salanti G. A two-stage prediction model for heterogeneous effects of treatments. *Stat Med*. 2021;40:4362-4375.
22. Steyerberg EW, Moons KGM, Windt DA, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10:e1001381.
23. Saramago P, Sutton AJ, Cooper NJ, Manca A. Mixed treatment comparisons using aggregate and individual participant level data. *Stat Med*. 2012;31:3516-3536.
24. Ghasemi N, Razavi S, Nikzad E. Multiple sclerosis: pathogenesis, symptoms, diagnoses and cell-based therapy. *Cell J Yakh-teh*. 2017;19:1-10.
25. Goldenberg MM. Multiple sclerosis review. *Pharm Ther*. 2012;37:175-184.
26. Rafiee Zadeh A, Askari M, Azadani NN, et al. Mechanism and adverse effects of multiple sclerosis drugs: a review article. Part 1. *Int J Physiol Pathophysiol Pharmacol*. 2019;11:95-104.
27. Hoepner R, Faissner S, Salmen A, Gold R, Chan A. Efficacy and side effects of natalizumab therapy in patients with multiple sclerosis. *J Cent Nerv Syst Dis*. 2014;6:41-49.
28. Disanto G, Benkert P, Lorscheider J, et al. The Swiss Multiple Sclerosis Cohort-study (SMSC): a prospective swiss wide investigation of key phases in disease evolution and new treatment options. *PLoS One*. 2016;11:e0152347.
29. Polman CH, O'Connor PW, Havrdova E, et al. A randomized, placebo-controlled trial of natalizumab for relapsing multiple sclerosis. *N Engl J Med*. 2006;354:899-910.
30. Gold R, Kappos L, Arnold DL, et al. Placebo-controlled phase 3 study of oral BG-12 for relapsing multiple sclerosis. *N Engl J Med*. 2012;367:1098-1107.
31. Fox RJ, Miller DH, Phillips JT, et al. Placebo-controlled phase 3 study of oral BG-12 or glatiramer in multiple sclerosis. *N Engl J Med*. 2012;367:1087-1097.
32. R Core Team. *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. 2019. Accessed October 7, 2023. <https://www.R-project.org/>
33. Plummer M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), March 20–22, Vienna Austria.
34. Harrell FE. *Regression Modelling Strategies: with Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer; 2015.
35. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer Science & Business Media; 2008.
36. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ*. 2009;338:b606.
37. Chalkou K, Steyerberg E, Bossuyt P, et al. Development, validation and clinical usefulness of a prognostic model for relapse in relapsing-remitting multiple sclerosis. *Diagn Progn Res*. 2021;5:17.
38. 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.
39. Sutton AJ, Cooper NJ, Jones DR. Evidence synthesis as the key to more coherent and efficient research. *BMC Med Res Methodol*. 2009;9:29.
40. Nordon C, Karcher H, Groenwold RHH, et al. The “efficacy-effectiveness gap”: historical background and current conceptualization. *Value Health*. 2016;19:75-81.
41. Ankarfeldt MZ, Adalsteinsson E, Groenwold RH, Ali MS, Klungel O. A systematic literature review on the efficacy-effectiveness gap: comparison of randomized controlled trials and observational studies of glucose-lowering drugs. *Clin Epidemiol*. 2017;9:41-51.

42. Steyerberg EW. Updating for a new setting. In: Steyerberg EW, ed. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer; 2008:361-389.
43. van Klaveren D, Balan TA, Steyerberg EW, Kent DM. Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. *J Clin Epidemiol*. 2019;114:72-83.
44. Hemming K, Hutton JL, Maguire MG, Marson AG. Meta-regression with partial information on summary trial or patient characteristics. *Stat Med*. 2010;29:1312-1324.
45. Riley RD, Steyerberg EW. Meta-analysis of a binary outcome using individual participant data and aggregate data. *Res Synth Methods*. 2010;1:2-19.
46. Sormani MP, Rovaris M, Comi G, Filippi M. A composite score to predict short-term disease activity in patients with relapsing-remitting MS. *Neurology*. 2007;69:1230-1235.
47. Held U, Heigenhauser L, Shang C, Kappos L, Polman C. Predictors of relapse rate in MS clinical trials. *Neurology*. 2005;65:1769-1773.
48. Liguori M, Meier DS, Hildenbrand P, et al. One year activity on subtraction MRI predicts subsequent 4 year activity and progression in multiple sclerosis. *J Neurol Neurosurg Psychiatry*. 2011;82:1125-1131.
49. Stühler E, Braune S, Lionetto F, et al. Framework for personalized prediction of treatment response in relapsing remitting multiple sclerosis. *BMC Med Res Methodol*. 2020;20:24. doi:10.1186/s12874-020-0906-6
50. Pellegrini F, Copetti M, Bovis F, et al. A proof-of-concept application of a novel scoring approach for personalized medicine in multiple sclerosis. *Mult Scler*. 2019;26(9):1064-1073.
51. Kalincik T, Manouchehrinia A, Sobisek L, et al. Towards personalized therapy for multiple sclerosis: prediction of individual treatment response. *Brain J Neurol*. 2017;140:2426-2443.
52. O'Hara RB, Sillanpää MJ. A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal*. 2009;4:85-117.
53. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons; 1987.
54. Carpenter JR, Kenward MG. *Multiple Imputation and Its Application*. John Wiley & Sons, Ltd; 2012.
55. Kent DM, Selker HP, Ruthazer R, Bluhmki E, Hacke W. The stroke-thrombolytic predictive instrument: a predictive instrument for intravenous thrombolysis in acute ischemic stroke. *Stroke*. 2006;37(12):2957-2962.
56. Deeks JJ, Higgins JPT, Altman DG. Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, et al., eds. *Cochrane Handbook for Systematic Reviews of Interventions version 6.4*. Cochrane Training; 2023. Accessed January 22, 2024. <https://training.cochrane.org/handbook/current/chapter-10>
57. Hoogland J, Int'Hout J, Belias M, et al. A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. *Stat Med*. 2021;40:5961-5981.
58. van Walraven C. Individual patient meta-analysis – rewards and challenges. *J Clin Epidemiol*. 2010;63:235-237.
59. Sud S, Douketis J. The devil is in the details... Or not? A primer on individual patient data meta-analysis. *Evid Based Med*. 2009;14:100-101.
60. Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Eval Health Prof*. 2002;25:76-97.
61. Buuren S v. *Flexible imputation of missing data*. Accessed November 25, 2021. <https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1893427> 2018.
62. Brown FS, Glasmacher SA, Kearns PKA, et al. Systematic review of prediction models in relapsing remitting multiple sclerosis. *PLoS One*. 2020;15:e0233575.
63. Van Calster B, Wynants L, Verbeek JFM, et al. Reporting and interpreting decision curve analysis: a guide for investigators. *Eur Urol*. 2018;74:796-804.
64. Antman EM, Cohen M, Bernink PJ, et al. The TIMI risk score for unstable angina/non-ST elevation MI: a method for prognostication and therapeutic decision making. *Jama*. 2000;284:835-842.
65. Morrow DA, Antman EM, Snapinn SM, McCabe C, Theroux P, Braunwald E. An integrated clinical approach to predicting the benefit of tirofiban in non-ST elevation acute coronary syndromes. Application of the TIMI risk score for UA/NSTEMI in PRISM-PLUS. *Eur Heart J*. 2002;23:223-229.
66. Cannon CP, Weintraub WS, Demopoulos LA, et al. Comparison of early invasive and conservative strategies in patients with unstable coronary syndromes treated with the glycoprotein IIb/IIIa inhibitor tirofiban. *N Engl J Med*. 2001;344:1879-1887.
67. Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. *Jama*. 2001;285:2864-2870.
68. Gage BF, van Walraven C, Pearce L, et al. Selecting patients with atrial fibrillation for anticoagulation: stroke risk stratification in patients taking aspirin. *Circulation*. 2004;110:2287-2292.
69. Vickers AJ, Calster BV, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352:i6.
70. van Klaveren D, Steyerberg EW, Serruys PW, Kent DM. The proposed 'concordance-statistic for benefit' provided a useful metric when modeling heterogeneous treatment effects. *J Clin Epidemiol*. 2018;94:59-68.
71. Efthimiou O, Hoogland J, Debray TPA, et al. Measuring the performance of prediction models to personalize treatment choice. *Stat Med*. 2023;42:1188-1206.
72. Keogh RH, van Geloven N. 2024 Prediction under interventions: evaluation of counterfactual performance using longitudinal observational data. <http://arxiv.org/abs/2304.10005>
73. Chalkou K, Vickers AJ, Pellegrini F, Manca A, Salanti G. Decision curve analysis for personalized treatment choice between multiple options. *Med Decis Making*. 2023, 3;43:337-349. doi:10.48550/arXiv.2202.02102
74. Heinze G, Boulesteix A-L, Kammer M, Morris TP, White IR, the Simulation Panel of the STRATOS initiative. Phases of methodological research in biostatistics—building the evidence base for new methods. *Biom J*. 2024;66:e2200222.
75. Verde PE, Ohmann C, Morbach S, Icks A. Bayesian evidence synthesis for exploring generalizability of treatment effects: a case study of combining randomized and non-randomized results in diabetes. *Stat Med*. 2016;35:1654-1675.



76. Hamza T, Chalkou K, Pellegrini F, et al. Synthesizing cross-design evidence and cross-format data using network meta-regression. *Res Synth Methods*. 2023;14:283-300. doi:[10.1002/jrsm.1619](https://doi.org/10.1002/jrsm.1619)

### SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Chalkou K, Hamza T, Benkert P, et al. Combining randomized and non-randomized data to predict heterogeneous effects of competing treatments. *Res Syn Meth*. 2024;1-16. doi:[10.1002/jrsm.1717](https://doi.org/10.1002/jrsm.1717)