Received: 10 May 2022 | Revised: 23 January 2023 | Accepted: 30 January 2023

DOI: 10.1002/jrsm.1625

RESEARCH ARTICLE

Research Synthesis Methods WILEY

Developing prediction models when there are systematically missing predictors in individual patient data meta-analysis

Michael Seo[1,2] | Toshi A. Furukawa[3] | Eirini Karyotaki[4,5,6] | Orestis Efthimiou[1,7]

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

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

[3]Departments of Health Promotion and Human Behavior and of Clinical Epidemiology, Kyoto University Graduate School of Medicine/School of Public Health, Kyoto, Japan

[4]Department of Global Health and Social Medicine, Harvard Medical School, Boston, USA

[5]Department of Clinical Neuro- and Developmental Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[6]Amsterdam Public Health Research Institute, Amsterdam, The Netherlands

[7]Department of Psychiatry, University of Oxford, Oxford, UK

Correspondence

Michael Seo, Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland.
Email: swj8874@gmail.com

Funding information

Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: Ambizione grant 180083

Abstract

Clinical prediction models are widely used in modern clinical practice. Such models are often developed using individual patient data (IPD) from a single study, but often there are IPD available from multiple studies. This allows using meta-analytical methods for developing prediction models, increasing power and precision. Different studies, however, often measure different sets of predictors, which may result to systematically missing predictors, that is, when not all studies collect all predictors of interest. This situation poses challenges in model development. We hereby describe various approaches that can be used to develop prediction models for continuous outcomes in such situations. We compare four approaches: a “restrict predictors” approach, where the model is developed using only predictors measured in all studies; a multiple imputation approach that ignores study-level clustering; a multiple imputation approach that accounts for study-level clustering; and a new approach that develops a prediction model in each study separately using all predictors reported, and then synthesizes all predictions in a multi-study ensemble. We explore in simulations the performance of all approaches under various scenarios. We find that imputation methods and our new method outperform the restrict predictors approach. In several scenarios, our method outperformed imputation methods, especially for few studies, when predictor effects were small, and in case of large heterogeneity. We use a real dataset of 12 trials in psychotherapies for depression to illustrate all methods in practice, and we provide code in R.

KEYWORDS

ensemble predictive modeling, individual patient data, meta-analysis, multilevel model, prediction research

Highlights

What is already known?

- Standard multiple imputation techniques can handle sporadically missing data, but they are not ideal for imputing systematically missing predictors.
- Several methods for imputing systematically missing variables while accounting for study clustering have been recently proposed.

What is new?

- We explore an alternative, generic method for addressing the systematically missing predictors' problem when the aim is to build a prediction model for a continuous outcome, using data from multiple studies.
- Instead of imputing systematically missing predictors, we propose to develop a separate prediction model in each study using only the predictors reported in that study, that is, ignoring systematically missing variables.

Potential impact for RSM readers outside the authors' field

- Imputation methods and ensemble method allow us to include additional predictors in our models, potentially increasing performance, or providing additional insight.
- We think that the ensemble method offers a potentially powerful alternative to researchers, and that it might be especially useful in the common case of having IPD from only a handful of studies, reporting different sets of predictors.

1 | INTRODUCTION

Meta-analysis of individual patient data (IPD) is often used to synthesize patient-level data from multiple studies, when developing clinical prediction models[1,2] or when estimating relative treatment effects.[3,4] One practical problem that often comes up in such analyses is missing data. This is the case when we have predictors in some studies only partly reported, for example, when some patients did not provide information on their BMI. We refer to this situation as 'sporadically missing' data. A different missing data problem, which is relevant for meta-analysis only, is 'systematically missing' predictors. This is when different studies measured different sets of predictors. Systematically missing predictors may pose practical problems in synthesizing data from multiple studies, for example, when we want to fit the same model across studies and meta-analyze the corresponding coefficients.

The most popular method for handling both types of missing data is multiple imputation.[5,6] Standard multiple imputation techniques[6] can handle sporadically missing data, but they are not ideal for imputing systematically missing predictors. When analyzing IPD from multiple studies, multiple imputation should ideally take into account the multilevel structure (i.e., clustering of patients in the different studies) and allow for potential heterogeneity between the studies. Several methods for imputing systematically missing variables while accounting for study clustering have been recently proposed.[7–9] However, applying these methods in practice may be difficult, because they require estimation of the variance–covariance matrix of random effects. This estimation uses studies with no systematically missing predictors, meaning that for valid inference, we need datasets with a large number of such studies[10]; of course, this may not always be the case.

This paper explores an alternative, generic method for addressing the systematically missing predictors' problem when the aim is to use data from multiple studies in order to build a model to predict a continuous outcome in future individuals who will have all covariates observed. Instead of imputing systematically missing predictors, we propose to develop a separate prediction model in each study using only the predictors reported in that study, that is, ignoring systematically missing variables. Then, we use these developed models for a new patient to make separate predictions, assuming that all covariates are collected for the new patient. In the end, we synthesize these predictions into a single forecast for the patient's future outcome.

We hereby compare in simulations the new approach with the restrict predictors approach where we exclude predictors that are systematically missing in some studies, the usual multiple imputation approach ignoring between-study heterogeneity and the multiple imputation approach where we account for between-study heterogeneity. Finally, we use a real dataset of trials in depression to illustrate the use of all methods in practice.

2 | ILLUSTRATIVE EXAMPLE IN DEPRESSION

We used a real dataset of 12 randomized trials in psychotherapies for depression comparing treatment as usual (TAU) versus internet-delivered cognitive behavioral therapy (iCBT). There were a total of 1633 patients randomized to TAU, 2072 to iCBT. The outcome of interest was Patient Health Questionnaire-9 scores (PHQ-9).[11] This is a measure of depression symptoms ranging from 0 to 27, with larger values indicating more severe depression. PHQ-9 was sporadically missing across all studies. There were also eight predictors of interest, which we will use to illustrate our methods. Among them, two were continuous (baseline PHQ-9 scores and age) and six binary (sex, relationship status, comorbid anxiety, previous episodes, medication, and alcohol use). Only baseline and sex were collected in all trials (albeit with few sporadically missing values for sex); the remaining predictors were collected inconsistently across trials. Table 1 shows the systematically missing data patterns across trials. In Table S1 of the Data S1, we show the percentages of sporadically missing data for each study and each outcome and predictor.

3 | METHODS

3.1 | Notation and general considerations

We assume that we have IPD from N_S studies. We assume that for patient i in study j, we have information on several predictors of interest, included in a vector $\boldsymbol{x_i}$, and that the observed outcome for this patient was y_i, measured on a continuous scale. We assume that different studies reported different sets of predictors, that is, different subsets from the full list of the predictors of interest. In case when patients in each study have been randomized to different treatments, we also have information on treatment t_i the patient received (either control $t_i = 0$, or active treatment $t_i = 1$).

Our aim is to build a model that will provide an accurate prediction of the outcome in new patients (possibly under different treatments). We will denote the predictors of a new patient as $\boldsymbol{x}_{\mathbf{new}}$, and a prediction of this patient outcome as $\widehat{y}_{\text{new}}$ (which will be a function of $\boldsymbol{x}_{\mathbf{new}}$, and maybe also of treatment). The focus will be on methods for handling systematically and sporadically missing predictors in the data. Of note, we will assume that all covariates will be collected for new patients.

3.2 | One-stage meta-analytical prediction models for fully observed data

In case all studies report all predictors and we have no sporadically missing data, we can fit a one-stage meta-

TABLE 1 Overview of the systematically missing data for each study in the illustrative example.

Study	Baseline depression score	Sex	Age	Relationship status	Comorbid anxiety	Previous episodes	Medication	Alcohol
De Graaf 2009	✓	✓	✓	✓	✓	✓	×	✓
Farrer 2011	✓	✓	✓	✓	✓	✓	×	✓
Geraedts 2014	✓	✓	✓	✓	✓	×	×	×
Gilbody 2015	✓	✓	×	✓	×	×	×	×
Johansson 2012	✓	✓	✓	✓	✓	×	✓	×
Kivi 2014	✓	✓	✓	✓	✓	✓	✓	✓
Klein 2016 (A)	✓	✓	✓	✓	✓	✓	✓	✓
Klein 2016 (B)	✓	✓	✓	✓	✓	✓	✓	✓
Meyer 2015	✓	✓	✓	✓	✓	✓	✓	×
Montero-Marin 2016	✓	✓	✓	✓	✓	×	×	×
Philips 2014	✓	✓	✓	✓	×	✓	✓	×
Rosso 2016	✓	✓	✓	×	✓	✓	✓	×

Note: ×, systematically missing in the study; ✓, available for at least some of the patients in the study.

analytical prediction model. We assume the (continuous) outcome to be normally distributed, $y_i \sim N(\mu_i, \sigma_i^2)$, where μ_i denotes the expected outcome of patient i (randomized in study j), and σ_i^2 refers to the residual variance of the outcome in each study. μ_i can be any function of the predictors $\boldsymbol{x_i}$, that is, $\mu_i = f(\boldsymbol{x_i})$. In the presence of treatment as a predictor, this function also includes t_i. The simplest prediction model we can build is one that does not account for study assignment, it assumes $\sigma_i^2 = \sigma$ and μ_i only includes linear terms:

$$\mu_i = \alpha + \boldsymbol{\beta} \boldsymbol{x_i}, \tag{1}$$

where $\boldsymbol{\beta}$ denotes the regression coefficients of the predictors $\boldsymbol{x_i}$. We can build a more advanced model by accounting for study assignment after including random effects in the intercept:

$$\begin{aligned} \mu_i &= a_j + \boldsymbol{\beta} \boldsymbol{x_i} \\ a_j &\sim N(\alpha, \tau_\alpha^2) \end{aligned}. \tag{2}$$

In this formula, a_j is the study-specific intercept, assumed to be normally distributed across studies with a mean α and variance τ_α^2.

Often when meta-analyzing comparative studies, we include a treatment effect term and/or treatment-predictor interactions ("effect modifiers"). With these included, μ_i can be modeled as follows:

$$\begin{aligned} \mu_i &= a_j + \boldsymbol{\beta} \boldsymbol{x_i} + \boldsymbol{\gamma} \boldsymbol{x_i}\, t_i + d_j\, t_i \\ a_j &\sim N(\alpha, \tau_\alpha^2), d_j \sim N(\delta, \tau_\delta^2) \end{aligned}. \tag{3}$$

In this formula, $\boldsymbol{\gamma}$ denotes the coefficients for treatment-predictor interactions and d_j is the study-specific 'baseline' treatment effect (i.e., at $\boldsymbol{x_i} = \boldsymbol{0}$), assumed to be normally distributed across studies with average δ and variance τ_δ^2.

After we fit any of the models described above, we can predict the outcome for a new patient with predictors $\boldsymbol{x_{new}}$ using the estimated values for the parameters of the model. In case of model (2) we would have

$$\widehat{y}_{\text{new}} = \widehat{\alpha} + \widehat{\boldsymbol{\beta}} \boldsymbol{x_{new}}. \tag{4}$$

In the case of analyzing comparative studies using model (3), predicted outcome would also be a function of treatment:

$$\widehat{y}_{\text{new}} = \widehat{\alpha} + \widehat{\boldsymbol{\beta}} \boldsymbol{x_{new}} + \widehat{\boldsymbol{\gamma}} \boldsymbol{x_{new}}\, t + \widehat{\delta}\, t. \tag{5}$$

We can extend Equations (4) or (5) by including non-linear terms and interactions between predictors, or by assuming exchangeable coefficients for main effects and treatment-predictor interactions across studies when fitting Equations (2) or (3). Furthermore, all these models can be fit in a two-stage approach.[12,13] All models described in this section are straightforward to fit when all studies report all predictors. Below, we discuss methods for fitting the models when some of the predictors are both systematically and sporadically missing.

3.3 | Methods for addressing missing predictors

3.3.1 | Restrict predictors method

In this approach, we need all studies to report the same predictors (i.e., no systematically, but possibly sporadically missing data). When this is not the case, we drop from model development predictors that were not reported in one or more studies. Obviously, this approach can be wasteful, especially when the ignored predictors are important.[14] Nonetheless, this is sometimes done when analyzing IPD.[15] After removing systematically missing predictors, we need to impute the sporadically missing predictors. Ideally, this should be done via multi-level imputation techniques. This approach can impute sporadically missing data while accounting for between-study heterogeneity.[6] Following this, we create m fully observed datasets and we use Equation (2) or (3) to develop a model in each one separately. Next, we need to pool results from these m prediction models. There are two different ways to do this. One method is to pool regression coefficients using Rubin's rules[16] and make predictions using Equations (4) or (5). A second approach is the "pooled prediction" strategy,[17] where Rubin's rules are applied on the predictions themselves. More specifically, given a new patient's predictors, we use the model developed in the k-th imputed dataset $(k = 1, 2, \ldots m)$ to obtain prediction $\widehat{y}_{\text{new},k}$. Finally, we pool the m predictions by simply taking the average:

$$\widehat{y}_{\text{new}} = \frac{1}{m} \sum_{k=1}^{m} \widehat{y}_{\text{new},k}. \tag{6}$$

The difference between two approaches is likely to be minor in practice, but the pooled prediction strategy may be preferable when the m predictions differ significantly.[17] Moreover, the pooled prediction strategy can be easily applied to any type of model, irrespective of the number of predictors or the structure of the model. We hereby follow this approach for the rest of the paper.

3.3.2 | Imputing missing predictors ignoring between-study heterogeneity

Instead of dropping predictors as above, we can use multiple imputation methods to impute both sporadically as well as systematically missing predictors simultaneously and create m fully observed datasets which include all predictors for all studies. The easiest way to do the imputations is to combine data from the studies in a single dataset and impute ignoring stratification of patients in different studies. Once we have the imputed datasets, we can use again Equation (2) or (3) to develop the model in each imputed dataset separately and then pool results from the m prediction models using Equation (6). Note that we are performing meta-analysis first and then using Rubin's rule to pool results.[18] The advantage of this approach is that it is straightforward and computationally easy to perform. The disadvantage is that it ignores the fact that subjects were randomized in different studies and makes a (perhaps strong) assumption when imputing, that the associations between variables are homogeneous across studies.[19] As it has been previously shown, however, this may lead to biased estimates[20] or underestimation of standard errors.[8]

3.3.3 | Imputing systematically missing predictors accounting for between-study heterogeneity

We already mentioned that standard multi-level imputation techniques based on linear mixed effects model can impute sporadically missing data while taking into account between-study heterogeneity.[6] This approach, however, cannot impute systematically missing data, since it requires estimation of the variance–covariance matrix of the predictors within each study; when for a study a predictor is systematically missing, this method fails.[10] Recently, three new methods were proposed to perform multiple imputations for the case of systematically missing predictors, while properly accounting for stratification of patients in studies and also imputing sporadically missing data: the method by Resche-Rigon and White,[7] by Jolani et al.,[8] and by Quartagno and Carpenter.[9] Audigier et al.[10] compared these methods in simulations and found that the method by Quartagno and Carpenter[9] may give biased results, when the number of studies is small. The other two methods provided more robust estimates when the number of studies was small. Audigier et al.[10] further recommended the method by Resche-Rigon and White[7] for relatively large sample sizes in the included studies. We follow this approach for the remaining of this paper. This method is a fully conditional specification imputation model, where a conditional distribution is defined for each incomplete variable.[6] It is a frequentist method based on a two-stage estimator. One key aspect of this method is that it requires some of the studies in the dataset to have no systematically missing predictors, to be able to estimate the relationships between all predictors. Specifically, for each study without systematically missing data, the method estimates regression coefficients of the effect of predictors on the outcome, variance–covariance matrix of these regression coefficients, and residual variance of the outcome. Then, at the second stage, it performs a multivariate meta-analysis of these study-specific estimates. Ultimately, using random draws from the estimated distribution of these parameters, the model imputes a value for all, systematically or sporadically, missing predictors.

One advantage of this method over the alternative two methods mentioned above (by Jolani et al.,[8] and by Quartagno and Carpenter[9]) is the computational speed since this method utilizes two-stage estimator as opposed to the one-stage estimator used in the other two methods.[10] A disadvantage is that with limited number of observations per study and large number of predictors, this method is more prone to overfitting.[10] Another limitation of the method is that, as mentioned above, we need to have at least two studies with no systematically missing predictors.

After using this method to impute missing data, we follow the steps described above, that is, use Equation (2) or (3) to develop the model in each imputed dataset, and Equation (4) or (5) to make predictions; then, Equation (6) to combine predictions.

3.3.4 | Ensemble method

Instead of imputing systematically missing predictors, we here present an alternative approach. Specifically, we propose fitting a different model in each study in the dataset, using only the predictors reported in that study, after only imputing sporadically missing data. This circumvents the problem of having studies not measuring at all certain predictors, without having to impute them. More specifically, in each study we first impute sporadically missing data as usual, using common multiple imputation methods. Thus, we obtain m full datasets for each study j. Next, we use each imputed dataset to fit a model using only the predictors that were reported in that study. The set of predictors we use for each study may be different, as some studies may not report some of the predictors. This means that, in principle, we may fit a different model in each study.

Thus, for each study we create m imputed datasets, and we fit the corresponding model there; at the end, for a new patient we have $N_s \times m$ different predictions. Let us denote prediction obtained from the k-th imputed dataset in study j as $\widehat{y}_{\text{new},k}^{(j)}$. To obtain the final prediction, we need to first combine the m predictions obtained from study j into a single estimate, by just taking the average, that is, $\widehat{y}_{\text{new}}^{(j)} = \frac{1}{m}\sum_{k=1}^{m}\widehat{y}_{\text{new},k}^{(j)}$. The variance of this estimate is given by the usual formula in multiple imputation,[16,21] that is

$$\operatorname{var}\left(\widehat{y}_{\text{new}}^{(j)}\right) = \frac{1}{m}\sum_{k=1}^{m}\operatorname{var}\left(\widehat{y}_{\text{new},k}^{(j)}\right) + \frac{1+m}{m(m-1)}\sum_{k=1}^{m}\left(\widehat{y}_{\text{new},k}^{(j)} - \widehat{y}_{\text{new}}^{(j)}\right)^2, \qquad (7)$$

where $\operatorname{var}\left(\widehat{y}_{\text{new},k}^{(j)}\right) = \left(\widehat{\sigma}_k^{(j)}\right)^2 \left(\boldsymbol{x}_{\mathbf{new}}^{\boldsymbol{T}}\left(\boldsymbol{X}_{\boldsymbol{k}}^{(\boldsymbol{j})\boldsymbol{T}}\boldsymbol{X}_{\boldsymbol{k}}^{(\boldsymbol{j})}\right)^{-1}\boldsymbol{x}_{\mathbf{new}} + 1\right)$ is the variance of the prediction of the outcome for the new patient using the model developed in the k-th imputed dataset in study j. The estimate $\widehat{\sigma}_k^{(j)}$ refers to the residual standard error of the outcome from the k-th imputed dataset for study j. Similarly, $\boldsymbol{X}_{\boldsymbol{k}}^{(\boldsymbol{j})}$ refers to the matrix of covariates from the k-th imputed dataset for study j. $\boldsymbol{x}_{\mathbf{new}}$ refers to the matrix of covariates for the new patient. Finally, after having estimated $\widehat{y}_{\text{new}}^{(j)}$ and $\operatorname{var}\left(\widehat{y}_{\text{new}}^{(j)}\right)$ from each study j, we obtain our final prediction $\widehat{y}_{\text{new}}$ as the weighted average of the study-specific predictions:

$$\widehat{y}_{\text{new}} = \frac{\sum_{j=1}^{N_s} w_{\text{new}}^{(j)}\,\widehat{y}_{\text{new}}^{(j)}}{\sum_{j=1}^{N_s} w_{\text{new}}^{(j)}}, \qquad (8)$$

where $w_{\text{new}}^{(j)} = 1/\operatorname{var}\left(\widehat{y}_{\text{new}}^{(j)}\right)$. The advantage of this method is that it takes fully into account stratification of patients in different studies, without requiring the existence of studies with no systematically missing predictors, as method of Section 3.3.3.

3.4 | Measuring performance of meta-analytical prediction models

After developing a prediction model, we want to measure its predictive performance. This may guide model selection (i.e., which of the four approaches described above, or which type of model should we employ when predicting outcomes for new patients?) or be used to gauge the usefulness of a model (i.e., is the model accurate enough?). Generally, assessing model performance is done by comparing model predictions with observations in a testing dataset. A usual measure of agreement between the two (for continuous outcomes) is the mean squared error (MSE):

$$\text{MSE} = \frac{1}{N}\sum\nolimits_i\left(\widehat{y}_i - y_i\right)^2, \qquad (9)$$

where N is the total number of patients in the testing dataset, $\widehat{y}_i$ is the predicted and y_i is the observed outcome for each patient i. Another common measure is the coefficient of determination (R-squared), showing the percentage of variance explained by the model:

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}},$$

where $\text{SS}_{\text{tot}} = \sum_i (y_i - \bar{y})^2$ is the total sum of squares, $\bar{y}$ is the average observed outcome, and $\text{SS}_{\text{res}} = \sum_i (\widehat{y}_i - y_i)^2$ is the residual sum of squares.

As a testing set, we can use the full dataset, that is, the dataset that was also used to develop the model. This approach is usually called "internal validation". Of note, internal validation may be prone to overfitting and subsequently optimism.[22] Overfitting implies that the model will predict very well in the data it was developed, but fail to predict well for new subjects. In such cases, assessment of model performance in internal validation will be optimistic. Overfitting will be a problem particularly when sample sizes are small and models are complicated, that is, including many predictors, higher-order terms of predictors, interactions of predictors, and so forth. One way to obtain optimism-adjusted estimates of model performance is to use resampling methods, for example, bootstrapping; for more details see the book by Steyerberg.[22]

Another, potentially useful approach is the so-called "internal-external" validation method for assessing model performance.[23] More specifically, since we are operating on a meta-analytical level, we can use a type of internal-external cross-validation called leave-one-study-out cross-validation (LOSO-CV) method.[24] In this procedure, one study is left out of the data and the rest are used for model development. The fitted model is used to make predictions about patients of the left-out study. Next, we use these predictions and observed outcomes to measure performance. The procedure is repeated after cycling through all studies, and measures of performance are summarized. This method allows us to also assess the transportability of the model, that is, to obtain an insight on how well it is expected to perform to other populations and settings.[2]

Systematically missing predictors, however, complicate both the internal validation and the LOSO-CV. The key point is that, to test a model in a dataset, we need to use patients with observations for all the predictors of the model, that is, without imputing missing values. The reason is that, if we impute, the apparent performance of the model would be affected by the accuracy of the imputation process, while we are only interested in assessing the capacity of the model to predict outcomes for new patients in clinical practice (where we can collect all the required predictors). For the case of the restrict predictors method, where the model is developed and tested using the same variables measured in all trials, systematically missing data poses no problems. This means that we can readily use internal validation and LOSO-CV without any complications to test the restrict predictors method in all studies, using only patients with fully observed predictors.

For the rest of the methods we described, however, things are not as straightforward. One simple way to address this issue is to limit testing to studies with no systematically missing data. This approach, however, might be problematic in practice if there are only few such studies. For the illustrative example, for internal validation we would be using all studies to develop the prediction models, but we would only be testing them in three, that is, Kivi 2014, Klein 2016 (A) and Klein 2016 (B) (see Table 1), in patients with no sporadically missing predictors. For the LOSO-CV we would exclude Kivi 2014, use the rest of the studies to develop the models, and then test in that study; then, we would do the same for the remaining two studies (Klein 2016 A and B), and summarize results.

If we do not have enough studies with no systematically missing data, or if we want to use data from all studies when comparing the competing strategies described above, we need to use an alternative approach, that is, we need to change the model we test in each study. More specifically, what we can do in order to test the two multiple imputation and ensemble approaches (Sections 3.3.2, 3.3.3, and 3.3.4) is use each study separately. For example, in the depression dataset, study De Graaf 2009 has a systematically missing predictor, history of medication (Table 1). In the two multiple imputation methods, all predictors can be used for imputing, but when developing the prediction model, we need to exclude history of medication. For internal validation we keep the data from De Graaf 2009 in the model-fitting process; for LOSO-CV, we take De Graaf 2009 out, and fit the model using the remaining studies. In both cases, after developing the model we test it using the patients in De Graaf 2009; finally, we cycle through all studies. In the ensemble method, we keep only the predictors that are not systematically missing in both the study we test, and the study we develop the model. For instance, in the internal validation, when we test in De Graaf 2009 we use all studies to develop models. When we build a model using the Geraedts 2014 study, we use only the five predictors that these two studies have measured in common (i.e., baseline, sex, age, relationship status, and comorbid-anxiety). For LOSO-CV the only difference is that De Graaf 2009 should be excluded from model fitting.

Obviously, this approach has drawbacks, most important being its complexity and the fact that the model we test may be different on each study we test it. However, it may serve as a rough guide for assessing the relative performance of the four approaches we described, when there are not enough studies with no systematically missing predictors.

4 | SIMULATION STUDY

4.1 | Overview of scenarios explored

We compared in simulations the performance of the four approaches described in the previous section for the prediction of a continuous outcome. We generated data under 64 different scenarios, where for each scenario we simulated 100 independent datasets. In these scenarios we explored various configurations regarding the number of studies, number of predictors, probability of predictors to be systematically missing in the studies, magnitude of the predictors' effects, and extent of heterogeneity. Since our focus was on methods for systematically missing predictors, and aiming to keep things relatively simple, we did not assume sporadically missing predictors. We explored the following configurations for the data-generating mechanisms:

1. Number of studies: 2, 3, 5 or 10.
2. Number of predictors: 5 or 10; of which 2 predictors were always reported, the rest might be systematically missing in each study.
3. Probability of systematically missing for each predictor in each study: 0.1 or 0.3.
4. Mean magnitude of the predictor effects on the outcome: 0.2 or 0.5.
5. Standard deviation of the magnitude of the predictor effects on the outcome across studies: 0.1 or 0.3.

Below we describe the procedure in more detail.

4.2 | Data generating mechanism

For each study in the dataset, we generated the number of patients by drawing from $U(150, 300)$. Predictors ($\boldsymbol{x}_i$)

were generated from a multivariate normal distribution with first-order autoregressive structure with homogeneous variances. The number of predictors was either 5 or 10 depending on the scenario. For 5 predictors, the predictors were generated from:

$$(x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}) \sim \mathrm{MVN}\left(\left(\chi_{j1}, \chi_{j2}, \chi_{j3}, \chi_{j4}, \chi_{j5}\right),\ \sigma_x^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix} \right),$$

where $\left(\chi_{j1}, \chi_{j2}, \chi_{j3}, \chi_{j4}, \chi_{j5}\right)$ are the overall study-specific means, σ_x^2 is the variance of the predictors across patients, ρ is the correlation between adjacent predictors. We sampled all $\boldsymbol{\chi_j}$ randomly for each scenario by drawing from $U(-1,1)$ and fixed $\rho = 0.2$, $\sigma_x^2 = 1$. Scenarios with 10 predictors were generated similarly. To simulate binary predictors, we categorized some of the generated continuous predictors. More specifically, in scenarios with 5 predictors, we made two of the predictors (x_{i2} and x_{i3}) binary, by categorizing at thresholds 0 and 0.5 respectively, for example, if $x_{i2} > 0$, we set $x_{i2} = 1$, 0 otherwise. In scenarios with 10 predictors, we made five of them binary (x_{i2}, x_{i3}, x_{i8}, x_{i9}, and x_{i10}), using thresholds at 0, 0, 0, 0.5, and 0.5 respectively.

The outcome was generated as $y_i = a_j + \boldsymbol{b_j x_i} + e_i$. a_j is the study intercept, randomly generated for each study by drawing from $U(0.5, 1.5)$. $\boldsymbol{b_j}$ is the vector of study-specific effects of the predictors, simulated as $b_{jk} \sim N\left(\mu_b, \sigma_b^2\right)$, where k refers to a distinct predictor, μ_b corresponds to the mean magnitude of the effect (0.2 or 0.5, according to the scenario) and σ_b corresponds to the standard deviation (0.1 or 0.3, according to the scenario). The random error was simulated for each patient separately as $e_i \sim N(0,1)$.

Finally, we assumed that the first 2 predictors (x_{i1} and x_{i2}) were always reported in all studies. The rest of predictors may be systematically missing in each study, with a fixed probability of missing. We explored scenarios where probability of missing was either 0.1 or 0.3, and we generated missing status by drawing from a Bernoulli distribution.

4.3 | Models compared and assessing predictive performance

After generating datasets as described above, we build prediction models following the four methods described in the previous section (i.e., restrict predictors method; multiple imputation without accounting for heterogeneity; multiple imputation accounting for heterogeneity; ensemble method). We used the model of Equation (2) assuming $\sigma_i^2 = \sigma$.

After developing all prediction models, we generated 10 new studies using the same data generating mechanism, to be used as testing data. The predictors in these studies were fully observed (no systematically missing predictors). We used the developed models and the data of the 10 new studies to make predictions, and we calculated MSE and R-squared for each model. In summary, for every scenario (64 in total) we generated 100 datasets. Each one of these 6400 datasets included 2, 3, 5 or 10 studies for developing the model (depending on the scenario), and 10 studies for testing the model.

As discussed, some scenarios had small number of studies (2, 3, 5), aiming to simulate realistic situations of data availability. In such scenarios, the multiple imputation method that accounted for heterogeneity might not be always feasible to implement. This is because, one of the limitations of this method is that it requires the existence of at least two studies with no systematically missing data, to be able to estimate variance covariance matrix; when this is not the case, this method fails. In order to make a fair comparison among the methods, we needed to compare them in the same datasets, that is, excluding datasets in which this method failed. At the same time, if for some scenarios this method failed very frequently, we would not have enough datasets to compare the rest of the models. Thus, we set a threshold of 20%. If for a specific scenario this method failed for less than 20% of the simulated datasets, we excluded these datasets from the analyses of all models. If it failed for more than 20%, we excluded this method from the analysis of this scenario. In addition, since we set a fixed probability that each predictor is systematically missing in each study, there might be datasets where one or more predictors were absent from all studies. In such datasets, all imputation methods will fail. Since such failures would be due to our data generating mechanism (and not due to a model's limitation), we excluded these datasets from all comparisons.

4.4 | Additional simulations on the effect of shrinkage

During the review process of our manuscript, one anonymous reviewer suggested exploring scenarios where the probability of a predictor to be reported in all studies (i.e., no systematically missingness) was related to the strength of the predictor. The same reviewer suggested to include in the simulations models that include shrinkage. Our new method, by omitting some covariates in some studies, does effectively perform a sort of shrinkage, so it is of interest to explore whether the possible advantages of this new method might be because of that, and whether these advantages would disappear if shrinkage methods were used. Thus, as additional exploratory analyses we added several scenarios in our simulations, where the predictor effects were different depending on their systematic missingness, and where we used a ridge regression model, combined with the restrict predictors and the multiple imputation methods.[25] More details are given in the Data S1.

4.5 | Implementation details

All analyses were carried out in R.[26] We used the **lme4** package[27] to fit linear mixed effects model. We used `glmnet` package to fit shrinkage models for the additional analysis. When we performed multiple imputations, the imputation model included all the predictors and the outcome.[28] As noted above, we did not generate sporadically missing predictors. Thus, for the restrict predictors and the ensemble method no imputation was required. For the method that ignores between-study heterogeneity, we used *pmm* method in the `mice`[7] package. This implements predictive mean matching based on the method by van Buuren.[6] For the imputation method accounting for between-study heterogeneity, we used *2l.2stage.norm* and *2l.2stage.bin* methods in the `micemd`[29] package depending on the type of the predictor to impute.[7] For all imputations, we imputed the missing variables to create $m = 10$ multiply imputed datasets. The R codes used for fitting all models are available at https://github.com/MikeJSeo/phd/tree/master/missing. Furthermore, the R package `bipd`, which is available in CRAN, implements all multiple imputation methods discussed in this article in a user-friendly manner. The vignette for the package demonstrates how to use this package in practice.

4.6 | Results

Table S3 in the Data S1 show the detailed simulation results. First, we found that—not surprisingly—the restrict predictors method that excludes systematically missing predictors was overall the worst approach, giving the largest MSE and smallest R-squared. Second, we found that for many datasets, the multiple imputation method that accounts for heterogeneity failed due to the unavailability of at least two studies with fully observed data. This was particularly frequent for scenarios with only 2 or 3 studies, and scenarios with larger number of predictors. Third, we found that the two multiple imputation methods performed best for both MSE and R-squared in scenarios with many studies (i.e., 10 studies), larger effect sizes of the predictors, and smaller heterogeneity of the effects. The difference between the two multiple imputation methods was trivial in most cases; however, the method that accounts for heterogeneity seemed to perform slightly better when the number of studies was large or when there was larger heterogeneity of the effects of the predictors. Conversely, we found that our new approach, the ensemble method, outperformed all other methods in scenarios with fewer studies (i.e., especially 2, 3, or even at 5 studies), when the effect of the predictors was relatively small, and when heterogeneity was relatively large, that is, at least a half of the mean predictor effect.

Table S4 of the Data S1 shows the additional simulations we performed, described in Section 4.4. The results again showed that even in situations where the effect of the predictor is related to its probability to be reported in the study, our new method usually outperformed other methods in scenarios where there was relatively large heterogeneity and small number of studies. Rather surprisingly, the new method performed better than other methods in scenarios when variables with systematically missingness had stronger effects than complete predictors (i.e., Scenarios R33–R48 in the Data S1), again especially for small number of studies. This might be due to the fact that performing wrong imputations can be detrimental when the predictor is strong, and using the ensemble method was preferable. When the reverse was true, that is, when complete predictors were stronger (scenarios R49–R64), results were not as clear. Again, for small number of studies and large heterogeneity our new method was usually best, however in many cases differences in performance were very small. In addition, we saw that shrinkage did not provide much benefit in all simulations. This might have been because all predictors had at least moderate effect in predicting the outcome. Overall, the results of the additional simulations did not affect our conclusions.

5 | ANALYSIS OF THE ILLUSTRATIVE EXAMPLE

5.1 | Implementation details

We used the data described in Section 2 to illustrate our methods. The data included a treatment indicator, several predictors, and a continuous outcome. We aimed at developing a model of the form $y_i \sim N(\mu_i, \sigma^2)$, that is, assuming common σ^2 across studies, and μ_i given by Equation (3). Since predictors were both sporadically and systematically missing, we used the four different methods described in the previous section to develop the prediction model.

Following the restrict predictors method (Section 3.3.1), we only included baseline PHQ-9 score and sex as predictors in the analysis, since these were the only predictors with no systematically missing values in all studies. There were however some few sporadically missing data on sex. We imputed these using MI, while accounting for the clustering of patients in different studies. To do this we used *2l.pmm* in the **miceadds**[30] package in R, which generalizes predictive mean matching via linear mixed models.[31] For the imputation method ignoring between-study heterogeneity, we used *pmm* method in **mice**[7] package and for the imputation method accounting for between-study heterogeneity, we used *2l.2stage.norm* or *2l.2stage.bin* method in **micemd**[29] depending on the type of the predictors to impute. For the ensemble method (Section 3.3.4), each study was used to develop an independent model, so multiple imputation did not need to be a multi-level procedure. Thus, we used *pmm* method in **mice**[7] package. For all imputations, we created $m = 20$ multiply imputed datasets. When imputing we used information from predictors, treatment, predictor-treatment interactions, and outcomes. Once multiply imputed datasets are created, we developed the prediction models after dropping patients with missing outcomes.

Next, we performed an internal and an internal-external (leave-one-study-out) cross validation of the modeling procedure. We did this following the two methods described in Section 3.4, that is, (A) using only studies with no systematically missing data; and (B) using all studies, after changing the model tested in each study as described in Section 3.4. For the internal validation we could not correct for optimism using bootstrapping, because some binary predictors were very rare in some of the studies (e.g., in one study there were only 2/301 patients reporting alcohol use; this means that bootstrapping sometimes resulted in samples with no patients on alcohol use). However, in this example we expected very low optimism, since overfitting was highly unlikely: the models included few predictors, the outcome was continuous, and the dataset was big. Finally note that for validation, we only used patients with complete data (i.e., no sporadically missing).

5.2 | Results

Following the restrict predictors method, we found baseline to be a strong predictor, but with weak evidence of

TABLE 2 Summary performance of the four different approaches for addressing missing data presented in this paper, using the depression dataset.

Validation method	Performance measure	Restrict predictors method	MI ignoring heterogeneity	MI accounting for heterogeneity	Ensemble method
Internal (A)	MSE	18.0	18.2	18.1	18.0
	R-squared	0.08	0.07	0.07	0.08
Internal–external LOSO-CV (A)	MSE	18.3	18.8	18.9	18.6
	R-squared	0.07	0.04	0.04	0.05
Internal (B)	MSE	26.1	26.1	26.2	26.2
	R-squared	0.18	0.19	0.18	0.18
Internal–external LOSO-CV (B)	MSE	26.6	26.9	26.9	26.7
	R-squared	0.17	0.16	0.16	0.17

Note: Approach (A) used only the three studies with no systematically missing data; and (B) used all studies, after changing the model tested in each study (details in Section 3.4).

Abbreviations: MI, multiple imputation; MSE, mean squared error.

an interaction with treatment. There was no evidence that sex could predict the outcome. Next, we analyzed the data following the two multiple imputation methods, including all predictors after imputation. Looking at the estimated parameters from the models, we see again baseline to be the most important predictor. For the remaining predictors estimates were more uncertain. Table S2 in the Data S1 show the estimated coefficients of the models. For the ensemble method, we report the simple average of estimates across different models developed for each study. We report zero for studies where the variable was systematically missing. Similarly, we simply average the variances of coefficient for each study.

In Table 2 we show the results from comparing the performance of the four different approaches using the internal cross validation and LOSO-CV. When we tested on the three studies with no systematically missing data, we saw that although all methods had similar performance, the restrict predictors method performed overall slightly better in terms of MSE and R-squared, for both internal and LOSO-CV. Among the more advanced methods, the ensemble method performed marginally better. When we tested using all studies, we saw that all four methods again had almost identical performance.

The reason why all methods led to similar results was most probably that in this example there was a dominant predictor (baseline severity) reported in all studies. Trying to include non-predictive variables brought in mostly noise, adding little benefit to the predictions. Additionally, one reason why the multiple imputation method that accounted for heterogeneity did not perform so well here might be the fact that this method performs multivariate meta-analysis using studies without systematically missing data. In this example, there were only three studies with no systematically missing predictors. This means that only three studies were used to estimate variance–covariance matrix of the random effects.

6 | DISCUSSION

This paper explored different methods for building prediction models when there are systematically missing predictors in individual patient data meta-analysis. Such models can then be used for predicting outcomes in future individuals with all covariates observed. We compared the performance of four methods (restrict predictors method, imputation method ignoring between-study heterogeneity, imputation method accounting for between-study heterogeneity, and our new ensemble-based approach). In the simulations, we investigated various scenarios for a different number of studies, number of predictors, probability of systematically missing studies for each predictor, and magnitude and heterogeneity of predictor effects. We found that the restrict predictors method was overall the worse approach. We also found that the ensemble method performed best for a few studies and when the systematically missing predictors have moderate-to-small effects with large heterogeneity. Conversely, for many studies, larger predictor effects and small heterogeneity, multiple imputation methods performed better. Among the two multiple imputation approaches, we found the one accounting for heterogeneity to be marginally better. Furthermore, we applied all models to a clinical example in psychotherapies for depression, where we saw only small differences between the four approaches. This was probably because in this example, there was a single strong predictor (baseline symptoms severity), consistently reported in all studies. In such cases, the choice between the various methods will be of minimal importance, as all methods will give practically the same results.

Our new method was inspired by the so-called ensemble learning models. This is a wide family of statistical and machine learning methods. In ensemble learning, multiple base models are combined to develop an overall prediction model. In many scenarios, this overall, ensemble model is expected to yield better predictive performance than each of its constituent parts.[32,33] Various combination techniques, such as the weighted average, can be employed for combining predictions from base-models. Despite their wide use, to the best of our knowledge, ensemble-based methods have not been previously used for addressing the problem of systematically missing data in IPD meta-analysis. In this work we aimed to fill this gap, by describing how an ensemble method can be employed when developing prediction models using data from multiple studies.

Several limitations of this work are worth mentioning. First, we did not explore the case of binary or time-to-event outcomes. This is an interesting area of future work, although there may be additional complications that need to be addressed.[34] Furthermore, in our simulations we only used simple data-generating mechanisms. Instead, we could have explored more complicated, and perhaps more realistic, mechanisms; for example, we could explore non-linear predictor-outcome associations when simulating data, or interactions between the predictors. Also, we could have included different types of statistical or machine learning prediction models, when assessing the performance of the methods for addressing the missing outcomes. For instance, although our simulation mechanism incorporated heterogeneity in coefficients for predictor effects, our analysis model used a linear model that assumed these coefficients to be common. Similar extensions could be pursued in future simulations. Furthermore, although we mentioned three alternative methods for imputing systematically missing variables while accounting for study-level clustering,[7-9] we only used the method by Resche-Rigon and White[7] in our simulations. However, we did some exploratory simulations using the method by Jolani,[8] but we did not find big differences with the method by Resche-Rigon and White; so we did not pursue this any further. Lastly, the biggest limitation of our new ensemble approach is perhaps the extra level of complexity it entails. Moreover, this approach requires building a separate model for each study and this can be time-consuming, especially when the number of studies is very large and when complex modeling strategies are employed. However, we think that for common situations of data availability this will not pose such a big problem.

In summary, in this paper we showed that more advanced methods may lead to better prediction models as compared to following the restrict predictors approach, in the presence of systematically missing data. These more advanced methods allow us to include additional predictors in our models, potentially increasing performance, or providing additional insight. In practice, we recommend researchers to select among the different methods after using both internal and internal-external cross-validation approaches. Finally, we think that the ensemble method offers a potentially powerful alternative to researchers, and that it might be especially useful in the common case of having IPD from only a handful of studies, reporting different sets of predictors.

AUTHOR CONTRIBUTIONS

Michael Seo programmed the analysis, performed the simulation study, and analyzed thedata. Michael Seo wrote the first draft of the manuscript and all authorscritically edited it. All authors approved thefinal submitted version of the manuscript.

ACKNOWLEDGMENTS

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: M.S. and O.E. were supported by the Swiss National Science Foundation (Ambizione grant number 180083). Open access funding provided by Universitat Bern.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created in this study.

ORCID

Michael Seo https://orcid.org/0000-0002-5229-590X
Toshi A. Furukawa https://orcid.org/0000-0003-2159-3776
Eirini Karyotaki https://orcid.org/0000-0002-0071-2599
Orestis Efthimiou https://orcid.org/0000-0002-0955-7572

REFERENCES

1. Steyerberg E, Nieboer D, Debray T, van Houwelingen H. Meta-analysis of prediction models. *Handbook of Meta-Analysis*. CRC Press; 2020. doi:10.1201/9781315119403-22
2. de Jong VMT, Moons KGM, Eijkemans MJC, Riley RD, Debray TPA. Developing more generalizable prediction models from pooled studies and large clustered data sets. *Stat Med*. 2021;40(15):3533-3559. doi:10.1002/sim.8981
3. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*. 2010; 340:c221. doi:10.1136/bmj.c221
4. 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(4):293-309. doi:10.1002/jrsm.1160
5. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc*. 1996;91(434):473-489. doi:10.2307/2291635
6. van Buuren S. *Flexible Imputation of Missing Data*. 2nd ed. Chapman and Hall/CRC; 2018.
7. Resche-Rigon M, White IR. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Stat Methods Med Res*. 2018;27(6):1634-1649. doi:10.1177/0962280216666564
8. Jolani S. Hierarchical imputation of systematically and sporadically missing data: an approximate Bayesian approach using chained equations. *Biom J*. 2018;60(2):333-351. doi:10.1002/bimj.201600220
9. Quartagno M, Carpenter JR. Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates. *Stat Med*. 2016;35(17):2938-2954. doi:10.1002/sim.6837
10. Audigier V, White I, Jolani S, et al. Multiple imputation for multilevel data with continuous and binary variables. *Stat Sci*. 2018;33(2):160-183. doi:10.1214/18-STS646
11. Furukawa TA, Suganuma A, Ostinelli EG, et al. Dismantling, optimising, and personalising internet cognitive behavioural therapy for depression: a systematic review and component network meta-analysis using individual participant data. *Lancet Psychiatry*. 2021;8(6):500-511. doi:10.1016/S2215-0366(21)00077-8
12. Debray TPA, Moons KGM, Abo-Zaid GMA, Koffijberg H, Riley RD. Individual participant data meta-analysis for a binary outcome: one-stage or two-stage? *PLoS One*. 2013;8(4):e60650. doi:10.1371/journal.pone.0060650
13. Riley RD, Debray TP, Fisher D, et al. Individual participant data meta-analysis to examine interactions between treatment effect and participant-level covariates: statistical recommendations for conduct and planning. *Stat Med*. 2020;39(15):2115-2137.
14. Jolani S, Debray TPA, Koffijberg H, van Buuren S, Moons KGM. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Stat Med*. 2015;34(11):1841-1863. doi:10.1002/sim.6451
15. Jackson D, White I, Kostis JB, et al. Systematically missing confounders in individual participant data meta-analysis of

observational cohort studies. *Stat Med*. 2009;28(8):1218-1237. doi:10.1002/sim.3540

16. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. John Wiley &Sons; 1987.
17. Wood AM, Royston P, White IR. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biom J*. 2015;57(4):614-632. doi:10.1002/bimj.201400004
18. Burgess S, White IR, Resche-Rigon M, Wood AM. Combining multiple imputation and meta-analysis with individual participant data. *Stat Med*. 2013;32(26):4499-4514. doi:10.1002/sim.5844
19. Resche-Rigon M, White IR, Bartlett JW, Peters SAE, Thompson SG, PROG-IMT Study Group. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Stat Med*. 2013;32(28):4890-4905. doi:10.1002/sim.5894
20. Reiter J, Raghunathan T, Kinney S. The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodol*. 2006;32(2), 143-150.
21. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med Res Methodol*. 2009;9(1):57. doi:10.1186/1471-2288-9-57
22. Steyerberg EW. *Clinical Prediction Models*. 2nd ed. Springer; 2019.
23. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal–external, and external validation. *J Clin Epidemiol*. 2016;69:245-247. doi:10.1016/j.jclinepi.2015.04.005
24. Debray TPA, Moons KGM, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med*. 2013;32(18):3158-3180. doi:10.1002/sim.5732
25. Hoerl A, Kennard R. Ridge regression: biased estimation for nonorthogonal problems. *Dent Tech*. 2012;12:55-67. doi:10.1080/00401706.1970.10488634
26. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2018. https://www.R-project.org/
27. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67(1):1-48. doi:10.18637/jss.v067.i01
28. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30(4):377-399. doi:10.1002/sim.4067
29. Audigier V, Resche-Rigon M. micemd: Multiple Imputation by Chained Equations with Multilevel Data; 2018. https://CRAN.R-project.org/package=micemd
30. Robitzsch A, Grund S. Miceadds: Some Additional Multiple Imputation Functions, Especially for "Mice"; 2021. https://CRAN.R-project.org/package=miceadds
31. Snijders T, Bosker R. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage; 1999.
32. Polikar R. Ensemble based systems in decision making. *IEEE Circuits Syst Mag*. 2006;6(3):21-45. doi:10.1109/MCAS.2006.1688199
33. Gashler M, Giraud-Carrier C, Martinez T. Decision tree ensemble: small heterogeneous is better than large homogeneous. *2008 Seventh International Conference on Machine Learning and Applications*; IEEE, 2008:900-905. doi:10.1109/ICMLA.2008.154
34. Pavlou M, Ambler G, Seaman S, Omar RZ. A note on obtaining correct marginal predictions from a random intercepts model for binary outcomes. *BMC Med Res Methodol*. 2015;15(1):1-6. doi:10.1186/s12874-015-0046-6

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: Seo M, Furukawa TA, Karyotaki E, Efthimiou O. Developing prediction models when there are systematically missing predictors in individual patient data meta-analysis. *Res Syn Meth*. 2023;14(3):455-467. doi:10.1002/jrsm.1625