Remiro-Azócar *et al.*
BMC Medical Research Methodology (2024) 24:32
https://doi.org/10.1186/s12874-024-02157-x

BMC Medical Research Methodology

RESEARCH

Open Access

Model-based standardization using multiple imputation

Antonio Remiro-Azócar[1*], Anna Heath[2,3,4] and Gianluca Baio[4]

Abstract

Background When studying the association between treatment and a clinical outcome, a parametric multivariable model of the conditional outcome expectation is often used to adjust for covariates. The treatment coefficient of the outcome model targets a conditional treatment effect. Model-based standardization is typically applied to average the model predictions over the target covariate distribution, and generate a covariate-adjusted estimate of the marginal treatment effect.

Methods The standard approach to model-based standardization involves maximum-likelihood estimation and use of the non-parametric bootstrap. We introduce a novel, general-purpose, model-based standardization method based on multiple imputation that is easily applicable when the outcome model is a generalized linear model. We term our proposed approach multiple imputation marginalization (MIM). MIM consists of two main stages: the generation of synthetic datasets and their analysis. MIM accommodates a Bayesian statistical framework, which naturally allows for the principled propagation of uncertainty, integrates the analysis into a probabilistic framework, and allows for the incorporation of prior evidence.

Results We conduct a simulation study to benchmark the finite-sample performance of MIM in conjunction with a parametric outcome model. The simulations provide proof-of-principle in scenarios with binary outcomes, continuous-valued covariates, a logistic outcome model and the marginal log odds ratio as the target effect measure. When parametric modeling assumptions hold, MIM yields unbiased estimation in the target covariate distribution, valid coverage rates, and similar precision and efficiency than the standard approach to model-based standardization.

Conclusion We demonstrate that multiple imputation can be used to marginalize over a target covariate distribution, providing appropriate inference with a correctly specified parametric outcome model and offering statistical performance comparable to that of the standard approach to model-based standardization.

Keywords Standardization, Marginalization, Multiple imputation, Parametric G-computation, Covariate adjustment, Indirect treatment comparisons

*Correspondence:
Antonio Remiro-Azócar
antonio.remiro-azocar@bayer.com
Full list of author information is available at the end of the article

Background

There has been active debate on whether marginal or conditional estimands should be preferred when estimating relative treatment effects [1–6]. Many researchers argue that marginal estimands are more appropriate inferential targets for decisions at the population level [1–3]. The distinction between marginal and conditional treatment effects is particularly important for non-collapsible measures such as the odds ratio and the hazard ratio. For such measures, the population-level marginal effect cannot be expressed as a weighted average of individual- or subgroup-level conditional effects. Almost invariably, marginal and conditional estimands do not coincide for non-collapsible effect measures, even in the absence of confounding and effect measure modification [7–12].

In the estimation of marginal treatment effects, covariate adjustment is desirable across a range of settings: (1) it is applied in the analysis of randomized controlled trials (RCTs) to correct for "chance" covariate imbalances and increase power, precision and efficiency [13]; (2) it allows for confounding control in the analysis of observational studies [14]; and (3) it accounts for covariate differences between multiple studies in indirect treatment comparisons and transportability analyses [15–19]. This article focuses on the latter scenario. Nevertheless, the methodological findings are also applicable to covariate adjustment between the treatment arms of a single comparative study.

A popular approach to covariate adjustment involves fitting a parametric multivariable model of the conditional outcome mean given treatment and baseline covariates. The treatment coefficient of the model targets a conditional effect. So-called model-based standardization, G-computation or marginalization approaches are required to integrate or average the conditional outcome model over the target covariate distribution, and produce a covariate-adjusted estimate of the marginal treatment effect [16, 20–29]. The standard approach to model-based standardization uses maximum-likelihood estimation to fit the outcome model and the non-parametric bootstrap for variance estimation [16, 25–28].

We introduce a novel general-purpose method for model-based standardization stemming from the ideas underlying multiple imputation [30]. Despite the close relationships between the methodologies, these largely have been developed separately, with some exceptions [31]. We build a link in this article,[1] terming our proposed approach *multiple imputation marginalization* (MIM). As opposed to the standard version of model-based standardization, MIM accommodates a Bayesian statistical framework, which naturally allows for the principled propagation of uncertainty, readily handles missingness in the patient-level data, integrates the analysis into a probabilistic framework, and permits the incorporation of prior evidence and other contextual information.

We conduct a simulation study to benchmark the finite-sample performance of MIM in conjunction with a parametric outcome model. The simulations provide proof-of-principle in scenarios with binary outcomes, continuous-valued covariates, a logistic outcome model and the marginal log odds ratio as the target effect measure. When parametric modeling assumptions hold, MIM yields unbiased estimation in the target covariate distribution. Code to implement the MIM methodology in `R` is provided in Additional file 1.

Methods

We wish to transport the results of a comparative "index" study to a target distribution of covariates. We assume that the target is characterized by a dataset that is external to the index study. In practice, this could belong to an observational study or be derived from secondary healthcare data sources (e.g. disease registries, cohort studies, insurance claims databases or electronic health records). Such administrative datasets are typically larger, less selected, and more representative of target populations of policy-interest than the participants recruited by RCTs [34–36].

For instance, in drug development, a pivotal Phase III RCT is typically conducted pre-market authorization to obtain regulatory approval. Such trial may have relatively narrow selection criteria, to enhance statistical precision and power in efficacy and safety testing [37–39]. Policymakers may be interested in transporting inferences to a "real-world" target covariate distribution, which is more diverse or heterogeneous in composition, and more representative of the patients who will receive the intervention in routine clinical practice [40].

Let $S = 1$ denote the index study and let $S = 2$ denote the external target. Adopting the potential outcomes framework [41], the target marginal average treatment effect estimand for the MIM procedure described in this article is a contrast between the, possibly transformed, means of potential outcome distributions:

$$TATE = g\Big(E\Big(Y^1 \mid S = 2\Big)\Big) - g\Big(E\Big(Y^0 \mid S = 2\Big)\Big), \tag{1}$$

where Y^t denotes the potential outcome that would have been observed for a subject assigned to intervention $T = t$, with $t \in \{0, 1\}$, $E(\cdot)$ represents an expectation taken over the distribution of potential outcomes in $S = 2$, and $g(\cdot)$ is an appropriate "link" function, e.g. the

[1] This article is based on research from Antonio Remiro-Azócar's PhD thesis [32] and a prior working paper by the authors [33].

identity, log or logit, mapping the mean potential outcomes onto the plus/minus infinity range. The target estimand in Eq. 1 is the average treatment effect, constructed on an additive scale e.g. the mean difference, (log) risk ratio or (log) odds ratio scale, had everyone in the target had been assigned $T = 1$ versus $T = 0$.

We briefly outline the data requirements of the MIM procedure. Individual-level data $\mathcal{D} = (\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{y})$ for a comparative index study, randomized or non-randomized, are available. Here, $\boldsymbol{x}$ is an $N \times K$ matrix of clinical or demographic baseline covariates, where N is the number of participants in the study and K is the number of baseline covariates. Each subject $n = 1, 2, \ldots, N$ has a row vector $\boldsymbol{x}_n = \left(x_{n,1}, x_{n,2}, \ldots, x_{n,K}\right)$ of K covariates. We let $\boldsymbol{y} = \left(y_1, y_2, \ldots, y_N\right)$ denote a vector of clinical outcomes and $\boldsymbol{t} = (t_1, t_2, \ldots, t_N)$ denote a binary treatment indicator vector, with each entry taking the value zero or one. We assume that $\mathcal{D}$ has no missing values but MIM can be readily adapted to address this issue, as is illustrated in Additional file 1.

The target dataset contains a matrix of covariates $\boldsymbol{x}^{tar}$ of dimensions $N^{tar} \times K$, where N^{tar} is the number of subjects and K is the number of covariates. We assume that all K covariates in the index study are available in the target. Each subject has a row vector $\boldsymbol{x}_i^{tar} = \left(x_{i,1}^{tar}, x_{i,2}^{tar}, \ldots, x_{i,K}^{tar}\right)$ of K covariates. Individual-level outcomes under the treatments being studied in the index trial are assumed unavailable in the target, as would be the case for interventions evaluated in the pre-marketing authorization setting.

In the scenario described in this article, standardization is performed with respect to an external data source, and the aim is to estimate marginal treatment effects in an external covariate distribution. This is typically the case in transportability analyses translating inferences from trials lacking external validity to the target population for decision-making, or in covariate-adjusted indirect comparisons transporting relative effects across a connected network of trials.

Nevertheless, as illustrated in Additional file 1, it is also possible to perform standardization over the covariate distribution observed in the index study. This avoids extrapolation into an external data source and may be useful when adjusting for covariate imbalances between treatment arms within randomized or non-randomized comparative studies. Within a randomized experiment, covariate adjustment is not necessary for unbiased estimation of the marginal treatment effect, but can be used to increase precision, i.e. reduce standard errors [13, 42]. Within a non-randomized study, covariate adjustment is necessary to remove confounding bias [43].

Multiple imputation marginalization

Conceptually, MIM consists of two separate stages: (1) the generation (*synthesis*) of synthetic datasets; and (2) the *analysis* of the generated datasets. The synthesis is separated from the analysis — only after the synthesis has been completed is the marginal effect of treatment on the outcome estimated. This is analogous to the separation between the imputation and analysis stages in multiple imputation.

Multiple imputation is a simulation technique that, arguably, is fundamentally Bayesian [30, 44, 45]. Its original development was grounded in Bayesian modeling, with imputed outcomes derived, at least conceptually, from a posterior predictive distribution. Computational tools such as Markov chain Monte Carlo (MCMC) and the Gibbs sampler only arose to prominence in the statistical literature several years after Rubin's seminal paper [46]. Consequently, the typical practical implementation of multiple imputation is based on a hybrid approach [30]: "think like a Bayesian, do like a frequentist".

Interestingly, our standardization problem can be conceptualized as a missing data problem. Outcomes for the subjects in the index study are observed, but outcomes in the target population, under the treatments examined in the index study, are systematically missing. MIM standardizes over the target by replacing the missing outcomes with a set of plausible values, conditional on some prespecified imputation mechanism. Extending the parallel with the missing data literature, MIM relies on a missing-at-random-like assumption: missing outcomes in the target are assumed conditionally exchangeable with those observed in the index study, conditioning on the adjustment model used for standardization.

MIM sits within a Bayesian framework by characterizing probabilistic relationships among a set of variables, and adopts a simulation approach. Figure 1 reveals a Bayesian directed acyclic graph (DAG) summarizing the general MIM structure and the links between its modules. In this graphical representation, the nodes represent variables; single arrows indicate probabilistic relationships and double arrows indicate deterministic functions. The plate notation indicates repeated analyses. We return to Fig. 1 and provide more detailed explanations for the notation and the individual modules throughout this section.

Generation of synthetic datasets: a missing data problem

The first stage, synthetic data generation, consists of two steps. Initially, the *first-stage regression* captures the relationship between the outcome $\boldsymbol{y}$ and the covariates $\boldsymbol{x}$ and treatment $\boldsymbol{t}$ in the patient-level data for the index study. In the *outcome prediction* step, predicted outcomes for

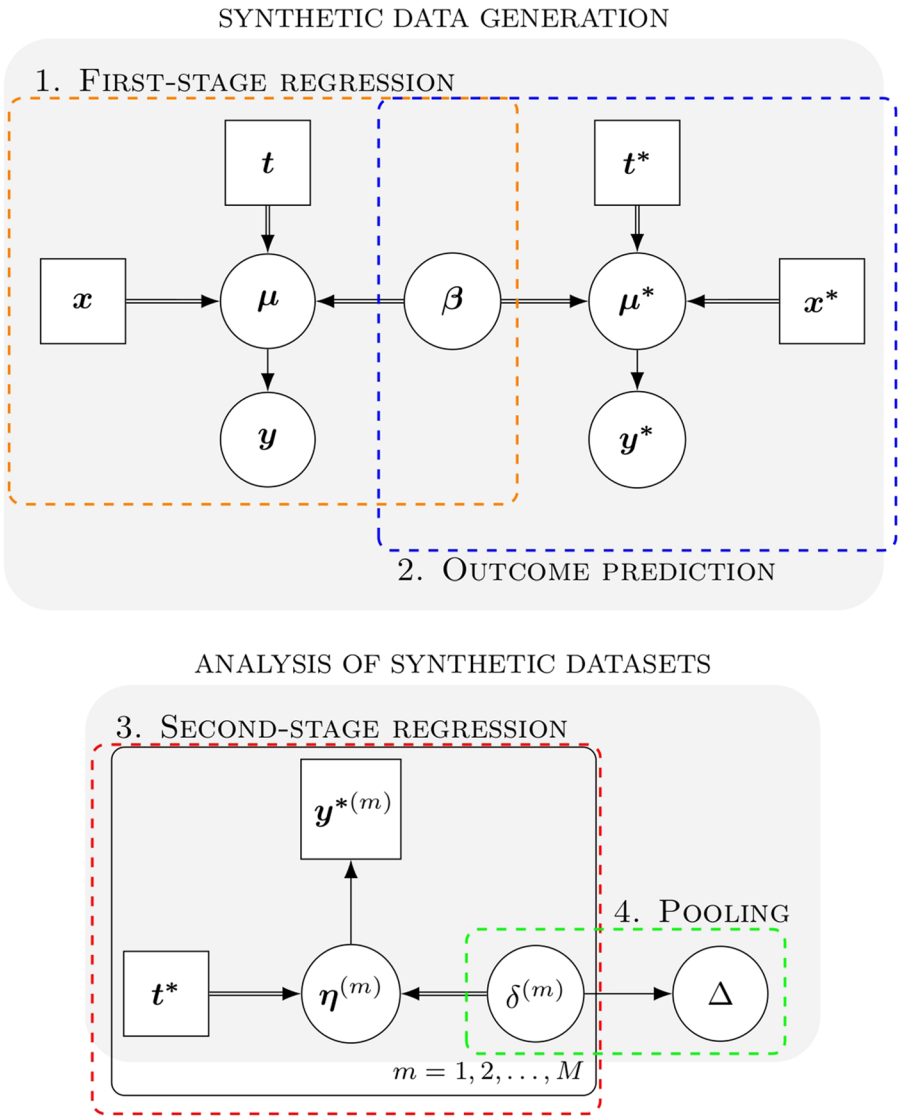


Fig. 1 Multiple imputation marginalization (MIM). A Bayesian directed acyclic graph representing MIM and its two main stages: (1) synthetic data generation; and (2) the analysis of synthetic datasets. Square nodes represent constant variables, circular nodes indicate stochastic variables, single arrows denote stochastic dependence, double arrows indicate deterministic relationships and the plate notation indicates repeated analyses

each treatment are generated in the target by drawing from the posterior predictive distribution of outcomes, given the observed predictor-outcome relationships in the index study, the set treatment and the target covariate distribution.

First-stage regression

Firstly, a multivariable regression of the observed outcome $\boldsymbol{y}$ on the baseline covariates $\boldsymbol{x}$ and treatment $\boldsymbol{t}$ is fitted to the subject-level data of the index study:

$$g(\mu_n) = \beta_0 + \boldsymbol{x}_n\boldsymbol{\beta_1} + \left(\beta_t + \boldsymbol{x}_n\boldsymbol{\beta_2}\right)1(t_n = 1), \qquad (2)$$

where μ_n is the conditional outcome expectation of subject n on the natural scale (e.g., the probability scale for binary outcomes), $g(\cdot)$ is an appropriate link function, β_0 is the intercept, $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are vectors of regression coefficients, and the treatment coefficient β_t targets a conditional effect at baseline, when the covariates are zero. The model specification assumes that the covariates are prognostic of outcome at the individual level. Due to the presence of treatment-covariate interactions, the covariates are also assumed to be (conditional) effect measure modifiers, i.e., predictive of treatment effect heterogeneity, at the individual level on the linear predictor scale.

The conditional outcome model in Eq. 2 will be our "working", "nuisance" or "imputation" model from now onward. We consider this to be a parametric model within a generalized linear modeling framework. In logistic regression, the link function $g(\mu_n) = \text{logit}(\mu_n) = \ln[\mu_n/(1-\mu_n)]$ is adopted. Other choices are possible in practice such as the identity link for linear regression or the log link for Poisson regression. The conditional outcome model is to be estimated using a Bayesian approach. We shall assume that efficient simulation-based sampling methods such as MCMC are used. Prior distributions for the regression coefficients would have to be specified, potentially using contextual information.

When standardizing with respect to an external target and/or when the index study is non-randomized, one is reliant on correct specification of the outcome model for unbiased estimation. In the former case, there is particular interest in modeling covariate-treatment product terms ("interactions") to capture (conditional) effect measure modification. In the latter case, the outcome model should adjust for potential confounders. Time and care should be dedicated to model-building, while being mindful of erroneous extrapolation outside the covariate space observed in the index study [47].

We shall assume that a single parametric outcome model is estimated, including treatment-covariate product terms to capture treatment effect heterogeneity at the individual level. An alternative strategy is to postulate two separate outcome models, one for each treatment group in the index comparative study [48]. While such approach allows for individual-level treatment effect heterogeneity over all the baseline covariates included in the models, it prevents borrowing information across treatment groups.

Outcome prediction

In this step, we generate predicted outcomes for the treatments under investigation, but in the target covariate distribution, by drawing from the posterior predictive distribution of outcomes. This is to be constructed using the imputation model in Eq. 2. Beforehand, a "data augmentation" step is required. We shall create a copy of the original target covariate dataset and vertically concatenate it to the original $\boldsymbol{x}^{tar}$. The concatenation is denoted $\boldsymbol{x}^* = \begin{bmatrix} \boldsymbol{x}^{tar} \\ \boldsymbol{x}^{tar} \end{bmatrix}$ and has $N^* = (2 \times N^{tar})$ rows and K columns. The original $j = 1, 2, \ldots, N^{tar}$ rows are assigned the treatment value $t_j^* = 1$ and the appended $j = (N^{tar}+1), (N^{tar}+2) \ldots, N^*$ rows are assigned the treatment value $t_j^* = 0$. The treatment indicator vector in the augmented dataset is denoted $\boldsymbol{t}^* = (t_1^*, t_2^* \ldots t_{N^*}^*)$.

Using MCMC sampling, it is fairly straightforward to implement the estimation of both the first-stage regression and the outcome prediction steps within a single Bayesian computation module. Having fitted the first-stage regression, we will iterate over the L converged draws of the MCMC algorithm to generate $M \leq L$ synthetic datasets: $\{\mathcal{D}^* = \mathcal{D}^{*(m)} : m = 1, 2 \ldots, M\}$, where $\mathcal{D}^{*(m)} = \left(\boldsymbol{x}^*, \boldsymbol{t}^*, \boldsymbol{y}^{*(m)}\right)$. Covariates $\boldsymbol{x}^*$ and treatment $\boldsymbol{t}^*$ are fixed across all the synthetic datasets. In line with the multiple imputation framework, each synthetic dataset is filled in by drawing a vector of outcomes $\boldsymbol{y}^{*(m)} = \left(y_1^{*(m)}, y_2^{*(m)}, \ldots, y_{N^*}^{*(m)}\right)$ of size N^* from the posterior predictive distribution $p(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{t}^*, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{t})$, given the original index trial and the augmented target datasets.

Assuming convergence of the MCMC sampling algorithm, the posterior predictive distribution of outcomes is approximated numerically as:

$$p(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{t}^*, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{t}) = \int_{\boldsymbol{\beta}} p(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{t}^*, \boldsymbol{\beta}) \cdot p(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{t}) d\boldsymbol{\beta}$$
$$\approx \frac{1}{L} \sum_{l=1}^{L} p\left(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{t}^*, \boldsymbol{\beta}^{(l)}\right),$$

where the realizations $\boldsymbol{\beta}^{(l)} \sim p(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{t})$ are independent draws from the posterior distribution of the first-stage regression parameters $\boldsymbol{\beta} = \left(\beta_0, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \beta_t\right)$, which encode the predictor-outcome relationships observed in the index trial, given some suitably defined prior $p(\boldsymbol{\beta})$. Here, $l = 1, 2, \ldots L$ indexes each MCMC iteration after convergence.

Consequently, L predictive samples $\boldsymbol{y}^{*(1)}, \ldots, \boldsymbol{y}^{*(L)} \sim p\left(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{t}^*, \boldsymbol{\beta}^{(l)}\right)$ are drawn independently from the posterior predictive distribution of outcomes. Dedicated MCMC programming software such as `Stan` [49] would typically return an $L \times N^*$ matrix of simulations. We will "thin" the matrix such that only $M \leq L$ of the rows are retained. This is to reduce the computational run time of the MIM analysis stage. The M remaining outcome imputations are used to complete the synthetic datasets $\{\mathcal{D}^{*(m)} = \left(\boldsymbol{x}^*, \boldsymbol{t}^*, \boldsymbol{y}^{*(m)}\right) : m = 1, 2, \ldots, M\}$. Table 1 illustrates the structure of each synthetic dataset.

Analysis of synthetic datasets

In the second stage, the analysis of synthetic datasets, we seek inferences about the marginal treatment effect in the target covariate distribution (*TATE* in Eq. 1, but here denoted Δ), given the synthesized outcomes. The analysis stage consists of another two steps. In the *second-stage regression* step, estimates of the marginal treatment effect in each synthesis $m = 1, 2, \ldots, M$ are generated by regressing the predicted outcomes $\boldsymbol{y}^{*(m)}$ on the treatment indicator $\boldsymbol{t}^*$. In the *pooling* step, the treatment effect

Table 1 An example of the structure of the m-th synthetic dataset, created in the data synthesis stage of MIM. In this example, $N = 7$ ($N^* = 14$) and $K = 3$. Prior to imputing the missing outcomes, a copy of the original target covariate dataset has been assigned the treatment value zero and been vertically concatenated to the original $\boldsymbol{x}^{tar}$, assigned the treatment value one

Covariates ($\boldsymbol{x}^*$)			Treatment ($\boldsymbol{t}^*$)	Outcome $\left(\boldsymbol{y}^{*(m)}\right)$
$x_{1,1}^{tar}$	$x_{1,2}^{tar}$	$x_{1,3}^{tar}$	1	$y_1^{*(m)}$
$x_{2,1}^{tar}$	$x_{2,2}^{tar}$	$x_{2,3}^{tar}$	1	$y_2^{*(m)}$
$x_{3,1}^{tar}$	$x_{3,2}^{tar}$	$x_{3,3}^{tar}$	1	$y_3^{*(m)}$
$x_{4,1}^{tar}$	$x_{4,2}^{tar}$	$x_{4,3}^{tar}$	1	$y_4^{*(m)}$
$x_{5,1}^{tar}$	$x_{5,2}^{tar}$	$x_{5,3}^{tar}$	1	$y_5^{*(m)}$
$x_{6,1}^{tar}$	$x_{6,2}^{tar}$	$x_{6,3}^{tar}$	1	$y_6^{*(m)}$
$x_{7,1}^{tar}$	$x_{7,2}^{tar}$	$x_{7,3}^{tar}$	1	$y_7^{*(m)}$
$x_{1,1}^{tar}$	$x_{1,2}^{tar}$	$x_{1,3}^{tar}$	0	$y_8^{*(m)}$
$x_{2,1}^{tar}$	$x_{2,2}^{tar}$	$x_{2,3}^{tar}$	0	$y_9^{*(m)}$
$x_{3,1}^{tar}$	$x_{3,2}^{tar}$	$x_{3,3}^{tar}$	0	$y_{10}^{*(m)}$
$x_{4,1}^{tar}$	$x_{4,2}^{tar}$	$x_{4,3}^{tar}$	0	$y_{11}^{*(m)}$
$x_{5,1}^{tar}$	$x_{5,2}^{tar}$	$x_{5,3}^{tar}$	0	$y_{12}^{*(m)}$
$x_{6,1}^{tar}$	$x_{6,2}^{tar}$	$x_{6,3}^{tar}$	0	$y_{13}^{*(m)}$
$x_{7,1}^{tar}$	$x_{7,2}^{tar}$	$x_{7,3}^{tar}$	0	$y_{14}^{*(m)}$

estimates and their variances are combined across all M syntheses.

In standard multiple imputation, the imputation and analysis stages may be performed simultaneously in a joint model [45]. In MIM, this is challenging because the dependent variable of the analysis is completely synthesized. Consider the Bayesian DAG in Fig. 1. In a joint model, the predicted outcomes are a collider variable, blocking the only path between the first and the second module (information from the directed arrows "collides" at the node). As a result, the data synthesis and analysis stages have been implemented as separate modules in a two-stage framework. The analysis stage conditions on the outcomes predicted by the synthesis stage, treating these as observed data.

Second-stage regression

We fit M second-stage regressions of predicted outcomes $\boldsymbol{y}^{*(m)}$ on treatment $\boldsymbol{t}^*$ for $m = 1, 2, \ldots, M$. Identical analyses are performed on each synthesis:

$$g\left(\eta_j^{(m)}\right) = \alpha^{(m)} + \delta^{(m)} t_j^*, \tag{3}$$

where $\eta_j^{(m)}$ is the expected outcome on the natural scale of unit j in the m-th synthesis, the coefficient $\alpha^{(m)}$ is an intercept term and $\delta^{(m)}$ denotes the marginal treatment effect in the m-th synthesis. There is some non-trivial computational complexity to performing a Bayesian fit in this step. That would embed a nested simulation scheme. Namely, if we draw M samples $\{\boldsymbol{y}^{*(m)} : m = 1, 2, \ldots M\}$ in the synthesis stage, a further number of samples, say R, of the treatment effect $\{\delta^{(m,r)} : m = 1, 2 \ldots M; r = 1, 2, \ldots R\}$ would be drawn for each of these realizations separately. This structure is unlikely to be feasible in terms of running time.

Using maximum-likelihood estimation, a point estimate $\hat{\delta}^{(m)}$ of the marginal treatment effect and a measure of its variance $\hat{v}^{(m)}$ are generated in each synthesis $\boldsymbol{y}^{*(m)}$. Equation 3 is a marginal model of outcome on treatment alone. Adopting terminology from the missing data literature, the second-stage regression in the analysis stage is "congenial" with the first-stage regression in the synthesis stage because treatment was already included as a predictor in the first-stage regression [44].

Pooling

We must now combine the M point estimates of the marginal treatment effect and their variances to generate a posterior distribution. Pooling across multiple syntheses is a topic that has already been investigated within the domain of statistical disclosure limitation [50–56].

In statistical disclosure limitation, data agencies mitigate the risk of identity disclosure by releasing multiple *fully synthetic* datasets. These only contain simulated values, in lieu of the original confidential data of real survey respondents. Raghunathan et al. [50] describe full synthesis as a two-step process: (1) construct multiple synthetic populations by repeatedly drawing from the posterior predictive distribution, conditional on a model fitted to the original data; and (2) draw random samples from each synthetic population, releasing these synthetic samples to the public. In practice, as indicated by Reiter and Raghunathan [55], it is not a requirement to generate the populations, but only to generate values for the synthetic samples. Once the samples are released, the analyst seeks inferences based on the synthetic data alone.

MIM is analogous to this problem, albeit there are some differences. In MIM, the analyst also acts as the synthesizer of data, and there is no "original data" on outcomes as such if the index study has not been conducted in the target covariate distribution. In any case, values for the samples are generated in the synthesis stage by repeatedly drawing from the posterior predictive distribution of outcomes. This is conditional on the predictor-outcome relationships indexed by the model fitted to the index study, the set treatment and the target distribution of covariates.

We seek to construct a posterior distribution for the marginal treatment effect, conditional on the synthetic outcomes (and treatment). That is, $p(\Delta \mid \boldsymbol{y}^*, \boldsymbol{t}^*)$. Following Raab et al. [56], each $\boldsymbol{y}^{*(m)}$ is viewed as a random sample from $p\left(\boldsymbol{y}^* \mid \boldsymbol{x}^*, \boldsymbol{t}^*, \boldsymbol{\beta}^{(m)}\right)$, where $\boldsymbol{\beta}^{(m)}$ is sampled from its posterior $p(\boldsymbol{\beta} \mid \boldsymbol{x}, \boldsymbol{t}, \boldsymbol{y})$. Hence, the "true" marginal treatment effect $\delta^{(m)}$ for the m-th synthesis, corresponding to $\boldsymbol{\beta}^{(m)}$, can be defined as a function of this sample. In each second-stage regression in Eq. 3, this is the treatment effect estimated by $\hat{\delta}^{(m)}$.

Consequently, following Raghunathan et al. [50], the estimators $\{\hat{\delta}^{(m)}, \hat{v}^{(m)}; m = 1, 2, \ldots, M\}$ from the second-stage regressions are treated as "data", and are used to construct an approximation to the posterior density $p(\Delta \mid \boldsymbol{y}^*, \boldsymbol{t}^*)$. This density is assumed to be approximately normal and is parametrized by its first two moments: the mean μ_Δ, and the variance σ^2_Δ. To derive the conditional distribution $p\left(\mu_\Delta, \sigma^2_\Delta \mid \boldsymbol{y}^*, \boldsymbol{t}^*\right)$ of these moments given the syntheses, the estimators $\{\hat{\delta}^{(m)}, \hat{v}^{(m)}; m = 1, 2, \ldots, M\}$, where $\hat{v}^{(m)}$ is the point estimate of the variance in the m-th second-stage regression, are treated as sufficient summaries of the syntheses, and μ_Δ and σ^2_Δ are treated as parameters. Then, the posterior distribution $p(\Delta \mid \boldsymbol{y}^*, \boldsymbol{t}^*)$ is constructed as:

$$p(\Delta \mid \boldsymbol{y}^*, \boldsymbol{t}^*) = \int_{\mu_\Delta, \sigma^2_\Delta} p\left(\Delta \mid \mu_\Delta, \sigma^2_\Delta\right) p\left(\mu_\Delta, \sigma^2_\Delta \mid \boldsymbol{y}^*, \boldsymbol{t}^*\right) d\left(\mu_\Delta, \sigma^2_\Delta\right). \tag{4}$$

In analogy with the theory of multiple imputation [30], the following quantities are required for inference:

$$\bar{\delta} = \sum_{m=1}^{M} \hat{\delta}^{(m)}/M, \tag{5}$$

$$\bar{v} = \sum_{m=1}^{M} \hat{v}^{(m)}/M, \tag{6}$$

$$b = \sum_{m=1}^{M} \left(\hat{\delta}^{(m)} - \bar{\delta}\right)^2/(M-1), \tag{7}$$

where $\bar{\delta}$ is the average of the treatment effect point estimates across the M syntheses, $\bar{v}$ is the average of the point estimates of the variance (the "within" variance), and b is the sample variance of the point estimates (the "between" variance). These quantities are computed using the point estimates from the second-stage regressions.

After deriving the quantities in Eqs. 5, 6 and 7, there are two options to approximate the posterior distribution of the marginal treatment effect in Eq. 4. The first involves direct Monte Carlo simulation and the second uses a simple normal approximation. In Additional file 1, the inferential framework for pooling outlined in this section is extended to scenarios involving correlated outcomes and non-scalar estimands with multiple components [57]. This involves a multivariate outcome model (i.e. with multiple dependent variables) and the combination of correlated treatment effects corresponding to multiple outcomes.

Pooling via posterior simulation Firstly, one draws μ_Δ and σ^2_Δ from their posterior distributions, conditional on the syntheses. These distributions are derived by Raghunathan et al. [50]. Values of μ_Δ are drawn from a normal distribution:

$$p\left(\mu_\Delta \mid \boldsymbol{y}^*, \boldsymbol{t}^*\right) \sim \mathrm{N}\left(\bar{\delta}, \bar{v}/M\right), \tag{8}$$

Values of σ^2_Δ are drawn from a chi-squared distribution with $M-1$ degrees of freedom:

$$p\left((M-1)b/\left(\sigma^2_\Delta + \bar{v}\right) \mid \boldsymbol{y}^*, \boldsymbol{t}^*\right) \sim \chi^2_{M-1}. \tag{9}$$

Values of Δ are drawn from a t-distribution with $M-1$ degrees of freedom [50]:

$$p\left(\Delta \mid \mu_\Delta, \sigma^2_\Delta\right) \sim t_{M-1}\left(\mu_\Delta, (1 + 1/M)\sigma^2_\Delta\right), \tag{10}$$

where the σ^2_Δ/M term in the variance is necessary as an adjustment for there being a finite number of syntheses; as $M \to \infty$, the variance tends to σ^2_Δ.

By performing a large number of simulations, one is estimating the posterior distribution in Eq. 4 by approximating the integral of the posterior in Eq. 10 with respect to the posteriors in Eqs. 8 and 9 [50]. Hence, the resulting draws of Δ are samples from the posterior distribution $p(\Delta \mid \boldsymbol{y}^*, \boldsymbol{t}^*)$ in Eq. 4. One can take the expectation over the posterior draws to produce a point estimate $\hat{\Delta}$ of the marginal treatment effect in the target distribution of covariates. A point estimate of its variance $\hat{V}\left(\hat{\Delta}\right)$ can be directly computed from the draws of the posterior density. Uncertainty measures such as 95% interval estimates can be calculated from the corresponding empirical quantiles.

The posterior distributions in Eqs. 8, 9 and 10 have been derived under certain normality assumptions, which are adequate for reasonably large sample sizes, where the relevant sample sizes are both the size N of the index study and the size N^* of the synthetic datasets.

Pooling via combining rules A simple alternative to direct Monte Carlo simulation is to use a basic approximation to the posterior density in Eq. 4, such that the

sampling distribution in Eq. 10 is normal as opposed to a t-distribution. The posterior mean is the average of the treatment effect point estimates across the M syntheses. A combining rule for the variance arises from using $b - \bar{v}$ to estimate σ^2_Δ, which is equivalent to setting σ^2_Δ at its approximate posterior mean in Eq. 9 [54]. Again, the b/M term is necessary as an adjustment for there being a finite number of syntheses.

Consequently, point estimates for the marginal treatment effect in the target covariate distribution and its variance can be derived using the following plug-in estimators:

$$\hat{\Delta} = \bar{\delta}, \tag{11}$$

$$\hat{V}\left(\hat{\Delta}\right) = (1 + 1/M)b - \bar{v}. \tag{12}$$

The combining rules are slightly different to Rubin's variance estimator in multiple imputation (in Eq. 12, $\bar{v}$ is subtracted instead of added) [30]. Interval estimates can be approximated using a normal distribution, e.g. for 95% interval estimates, taking ± 1.96 times the square root of the variance computed in Eq. 12 [50]. A more conservative, heavier-tailed, t-distribution with $\nu_f = (M-1)(1 + \bar{v}/((1 + 1/M)b))^2$ degrees of freedom has also been proposed, as normal distributions may produce excessively narrow intervals and undercoverage when M is more modest [52]. Note that the combining rules in Eqs. 11 and 12 are only appropriate for reasonably large M. The choice of M is now discussed.

Number of synthetic datasets

In standard multiple imputation, it is not uncommon to release as little as five imputed datasets [30]. However, MIM is likely to require a larger value of M because it imputes all of the outcomes in the syntheses, as opposed to a relatively small proportion of missing values. Adopting terminology from the missing data literature, the "fraction of missing information" in MIM is 1, because the original dataset used to fit the first-stage regression is different than the augmented target dataset used to fit the second-stage regression.

In the statistical disclosure limitation literature, a common choice for the number of syntheses is $M = 100$ [52]. We encourage setting M as large as possible, in order to minimize Monte Carlo error and thereby maximize precision and efficiency. A sensible strategy is to increase the number of syntheses until repeated analyses across different random seeds give similar results, within a specified degree of accuracy. Assuming MCMC simulation is used in the synthesis stage, the value of M is likely to be a fraction of the total number of iterations or posterior samples required for convergence. As computation time is driven by the synthesis stage, increasing M provides more precise and efficient estimation [52, 58] at little additional cost in the analysis stage.

An inconvenience of the expressions in Eqs. 9 and 12 is that these may produce negative variances. When the posterior in Eq. 9 generates a negative value of σ^2_Δ, i.e., when $\frac{(M-1)b}{\chi^*} < \nu$ (where χ^* is the draw from the posterior in Eq. 9), the variance of the posterior distribution in Eq. 10 is negative. Similarly, Eq. 12 produces a negative variance when $(1 + 1/M)b < \bar{v}$. This is because the formulations have been derived using method-of-moments approximations, where estimates are not necessarily constrained to fall in the parameter space. Negative variances are unlikely to occur if M and the size of the synthetic datasets are relatively large. This is due to lower variability in σ^2_Δ and $\hat{V}\left(\hat{\Delta}\right)$ [53]: $\bar{v}$ decreases with larger syntheses and b is less variable with larger M [52].

Simulation study

Aims

The objectives of the simulation study are to provide proof-of-principle for MIM and to benchmark its statistical performance against that of the standard implementation of parametric model-based standardization (parametric G-computation), which uses maximum-likelihood estimation with non-parametric bootstrapping for inference [16, 25–28]. The simulation study investigates a setting in which the index study is a perfectly-executed two-arm RCT. This will be standardized to produce a marginal treatment effect in an external target covariate distribution.

Methods will be evaluated according to the following finite-sample (frequentist) characteristics [59]: (1) unbiasedness; (2) precision; (3) efficiency; and (4) coverage of interval estimates. The chosen performance metrics assess these criteria specifically. The ADEMP (Aims, Data-generating mechanisms, Estimands, Methods, Performance measures) structure by Morris et al. [59] is used to describe the simulation study design. Example `R` code implementing the methods on a simulated example is provided in Additional file 1. All simulations and analyses have been performed using `R` software version 4.1.1 [60].[2]

Data-generating mechanisms

The data-generating mechanisms are inspired by those presented by Phillippo et al. [61]. We consider binary outcomes using the log odds ratio as the measure of effect. An index RCT investigates the efficacy of an active

[2] The files required to run the simulations are available at http://github.com/remiroazocar/MIM.

treatment (coded as treatment 1) versus a control (coded as treatment 0). Outcome y_n for subject n in the index RCT is simulated from a Bernoulli distribution, with event probabilities $\theta_n = p(y_n \mid \boldsymbol{x}_n, t_n)$ given covariates $\boldsymbol{x}_n$ and treatment t_n generated using a logistic model:

$$y_n \sim \text{Bernoulli}(\theta_n),$$
$$\theta_n = \text{logit}^{-1}\left[\beta_0 + \beta_{1,1}x_{n,1} + \beta_{1,2}x_{n,2} + \left(\beta_t + \beta_{2,1}x_{n,1} + \beta_{2,2}x_{n,2}\right)1(t_n = 1)\right].$$

Two correlated continuous covariates, $x_{n,1}$ and $x_{n,2}$ are simulated per subject n by drawing from a multivariate Gaussian copula with pre-specified means and standard deviations for the marginal distributions, and a pre-specified covariance matrix. The first covariate follows the marginal distribution $x_{n,1} \sim \text{N}(1, 0.5^2)$, and the second covariate follows the marginal distribution $x_{n,2} \sim \text{N}(0.5, 0.2^2)$. There is some positive correlation between the two covariates, with pairwise correlation coefficients set to 0.15. Both covariates are prognostic of the outcome in the control group at the individual level. Due to the presence of treatment-covariate interactions, both covariates are also (conditional) effect measure modifiers (i.e., predictive of treatment effect heterogeneity) at the individual level on the (log) odds ratio scale. The covariates also modify marginal treatment effects at the population level on the (log) odds ratio scale.

We set the intercept to $\beta_0 = -0.5$, coefficients for the main covariate-outcome associations to $\beta_{1,k} = 2\sigma_k$ for covariate k, where σ_k is the standard deviation of the sampling distribution of covariate k, and coefficients for the interaction terms to $\beta_{2,k} = \sigma_k$. The treatment coefficient (i.e. the conditional log odds ratio for the active intervention versus control at baseline, when the covariate values are zero) is set to $\beta_t = -1.5$. That is, if the binary outcome represents the occurrence of an adverse event, the active treatment would be more efficacious than the control. The covariates could represent influential prognostic and effect-modifying comorbidities that are associated with greater odds of the adverse event and lower efficacy of active treatment versus control at the individual level on the (log) odds ratio scale.

The simulation study adopts a factorial arrangement using three index trial sample sizes times two levels of overlap between the index trial and the target covariate distributions. This results in a total of six simulation scenarios. The settings are defined by varying the following parameter values:

- Sample sizes of $N \in \{500, 1000, 2000\}$ for the index RCT, with a 1:1 active treatment vs. control allocation ratio.
- The level of (deterministic) overlap between the index RCT and the target covariate distribution: limited overlap (50% of the index study population lies outside of the target population) and full overlap (the index study population is entirely contained within the target population) [61].

Following Phillippo et al. [61], the target covariate distribution is set to achieve the required level of overlap by using a proxy parameter κ ($\kappa = 0.5$ corresponds to 50% overlap and $\kappa = 1$ corresponds to full overlap). Then, each covariate k in the target follows the marginal distribution $x^*_{n,k} \sim \text{N}\left(m^*_k, \sigma^{*2}_k\right)$, with $m^*_k = m_k(1.1 + (1-\kappa)^2)$ and $\sigma^*_k = 0.75\sigma_k$, where m_k is the mean of the sampling distribution of covariate k in the index RCT. The target joint covariate distribution is a multivariate Gaussian copula with the pairwise correlation coefficients set to 0.15. $N^{tar} = 2000$ subject profiles are simulated for the target covariate dataset. Individual-level outcomes in the target, under the treatments being investigated in the index study, are assumed unavailable and not simulated.

Estimands

The target estimand is the true marginal log odds ratio for active treatment versus control in the target covariate distribution. This may vary across the settings of the simulation study because, by design, changing the level of (deterministic) overlap changes the target covariate distribution, and the true marginal log odds ratio depends on the covariate distribution.

For each scenario, true values of the marginal estimand are determined by simulating a cohort of 2,000,000 subjects, a number sufficiently large to minimize sampling variability, using the target covariate distributions in the simulation study. Hypothetical subject-level binary outcomes under active treatment and control are simulated for the cohort according to the true outcome-generating mechanism. The true marginal log odds ratio is computed by averaging the simulated unit-level outcomes under each treatment and contrasting the marginal outcome expectations on the log odds ratio scale. A simulation-based approach is necessary to compute the true marginal estimands due to the non-collapsibility of the (log) odds ratio [62–64].

For $\kappa = 0.5$ (limited overlap), the true marginal outcome probabilities for active treatment and control in the

target population are 0.60 and 0.75, respectively, resulting in a true marginal log odds ratio of -0.68. For $\kappa = 1$ (full overlap), the true marginal outcome probabilities for active treatment and control in the target population are 0.50 and 0.69, resulting in a true marginal log odds ratio of -0.81.[3]

Methods

Each simulated dataset is analyzed using: (1) the standard implementation of parametric model-based standardization (parametric G-computation) [16, 25–28]; and (2) parametric model-based standardization using MIM.

Standard model-based standardization

Among the subjects in the index RCT, outcomes are regressed on baseline covariates and treatment using a logistic model. Maximum-likelihood is used to estimate the conditional outcome model, which is correctly specified. Outcome predictions under each treatment are made by applying the fitted regression to the full target covariate dataset. The marginal log odds ratio is derived by: (1) averaging the predicted conditional outcome means by treatment group over the target covariate dataset; (2) transforming the resulting marginal outcome means to the log odds ratio scale; and (3) producing a contrast for active treatment versus control in such scale [16, 27, 28]. For inference, the index RCT is resampled via the ordinary non-parametric bootstrap with replacement, using 1,000 resamples (the target covariates are assumed fixed). The average marginal log odds ratio and its standard error are computed as the mean and the standard deviation, respectively, across the resamples. Confidence intervals are computed using the "percentile" method; 95% interval estimates are derived from the 2.5th and the 97.5th percentiles across the resamples.

Multiple imputation marginalization

In the synthesis stage, the first-stage multivariable logistic regression is correctly specified and is estimated using MCMC sampling. This is implemented using the `R` package `rstanarm` [65], an appendage to `rstan` [66]. We adopt the default normally-distributed "weakly informative" priors for the logistic regression coefficients [65]. Predicted outcomes are drawn from their posterior predictive distribution, given the augmented target dataset. We run two Markov chains with 4,000 iterations per chain, with 2,000 "burn-in" iterations that are not used for posterior inference. The MCMC chains are thinned every 4 iterations to use a total of $M = (2000 \times 2)/4 = 1000$ syntheses of size $N^* = 2 \times N^{tar} = 4000$ in the analysis stage. The second-stage regressions are simple logistic regressions of predicted outcomes on treatment that are fitted to each synthesis using maximum-likelihood estimation. Their point estimates and variances are pooled using the combining rules in Eqs. 11 and 12. Wald-type 95% confidence intervals are estimated using t-distributions with $\nu_f = (M-1)(1+\bar{\nu}/((1+1/M)b))^2$ degrees of freedom. Variance estimates are never negative under any simulation scenario. In a test simulation scenario ($\kappa = 0.5$ and $N = 1000$), the selected value of $M = 1000$ is high enough, so that the Monte Carlo error is adequate with respect to the uncertainty in the estimator. Upon inspection, marginal log odds ratio estimates across different random seeds are approximately within 0.01.

Performance measures

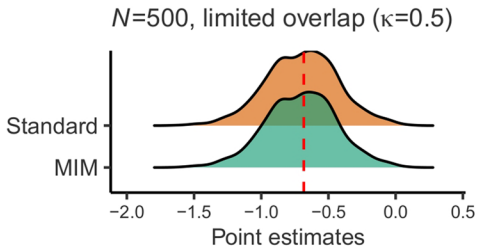
We simulate 1,000 datasets per scenario. For each scenario and methodology, the following performance measures are computed over the 1,000 simulated datasets: (1) bias; (2) empirical standard error (ESE); (3) mean square error (MSE); and (4) empirical coverage rate of the 95% interval estimates. These criteria are explicitly defined by Morris et al. [59] and specifically address the aims of the simulation study. The ESE evaluates precision (aim 2) and the MSE measures overall efficiency (aim 3), accounting for bias (aim 1) and precision (aim 2).

To quantify the simulation uncertainty, Monte Carlo standard errors (MCSEs) over the data replicates, as defined by Morris et al. [59], will be reported for each performance metric. Based on the scenario inducing the highest long-run variability ($\kappa = 0.5$ and $N = 500$), the MCSE of the bias of the methods is at most 0.015 under 1,000 simulations per scenario, and the MCSE of the coverage (based on an empirical coverage percentage of 95%) is $\left(\sqrt{(95 \times 5)/1000}\right)\% = 0.69\%$, with the worst-case being 1.58% under 50% coverage. Such levels of simulation uncertainty are considered sufficiently precise, and 1,000 simulations per scenario are deemed appropriate.

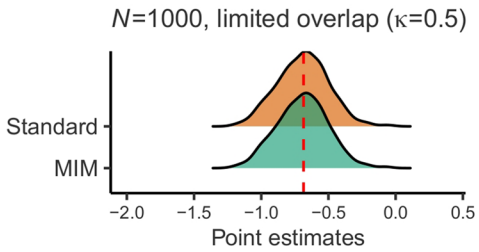
Results

Performance metrics for the six simulation scenarios are reported in Fig. 2. The limited overlap settings ($\kappa = 0.5$) are displayed at the top (in ascending order of index trial sample size, from top to bottom), followed by the full overlap settings ($\kappa = 1$) at the bottom. For each simulation scenario, there is a ridgeline plot depicting the spread of point estimates for the marginal log odds ratio across the 1,000 simulation runs. The dashed red lines indicate the true estimands. At the right of each ridgeline plot, a summary tabulation exhibits empirical quantities

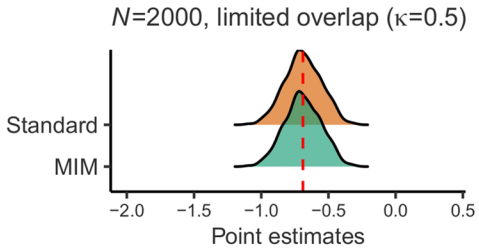
[3] In contrast, the true average conditional log odds ratio for active treatment versus control, at the covariate means of the target population, is given by the weighted average: $\beta_t + \beta_{2,1} m_1^* + \beta_{2,2} m_2^* = -2.5 + 0.5 \times (1.1 + (1-\kappa)^2) + 0.2 \times 0.5 \times (1.1 + (1-\kappa)^2)$. This is equal to -0.69 for $\kappa = 0.5$ and to -0.84 for $\kappa = 1$.



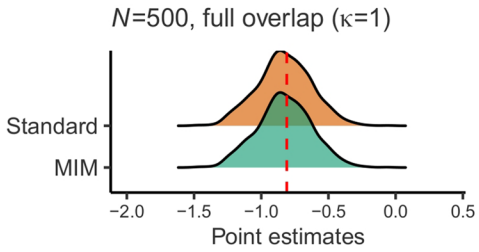
Method	Bias	ESE	MSE	Coverage
Standard	−0.013 (0.008)	0.267 (0.006)	0.071 (0.003)	0.947 (0.007)
MIM	−0.019 (0.008)	0.263 (0.006)	0.070 (0.003)	0.947 (0.007)



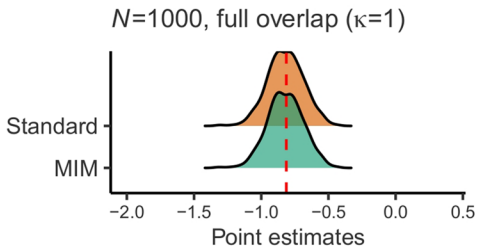
Method	Bias	ESE	MSE	Coverage
Standard	−0.004 (0.006)	0.192 (0.004)	0.037 (0.002)	0.952 (0.007)
MIM	−0.007 (0.006)	0.191 (0.004)	0.037 (0.002)	0.953 (0.007)



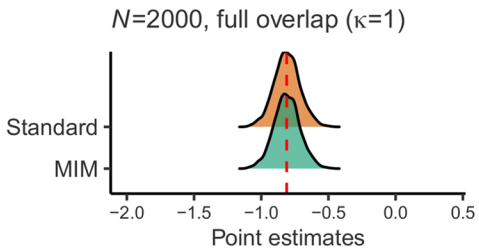
Method	Bias	ESE	MSE	Coverage
Standard	0.002 (0.004)	0.136 (0.003)	0.018 (0.001)	0.951 (0.007)
MIM	0.000 (0.004)	0.136 (0.003)	0.018 (0.001)	0.953 (0.007)



Method	Bias	ESE	MSE	Coverage
Standard	−0.017 (0.006)	0.203 (0.005)	0.042 (0.002)	0.935 (0.008)
MIM	−0.017 (0.006)	0.203 (0.005)	0.041 (0.002)	0.934 (0.008)



Method	Bias	ESE	MSE	Coverage
Standard	−0.006 (0.004)	0.134 (0.003)	0.018 (0.001)	0.958 (0.006)
MIM	−0.006 (0.004)	0.134 (0.003)	0.018 (0.001)	0.958 (0.006)



Method	Bias	ESE	MSE	Coverage
Standard	0.002 (0.003)	0.096 (0.002)	0.009 (0.000)	0.945 (0.007)
MIM	0.002 (0.003)	0.096 (0.002)	0.009 (0.000)	0.940 (0.008)

Fig. 2 Simulation study results. The distribution of treatment effect point estimates over the simulation runs and the empirical quantities used to measure the statistical performance of standard model-based standardization ("Standard") and multiple imputation marginalization ("MIM") are visualized for the six scenarios. The dashed red lines in the ridgeline plots to the left indicate the true estimands

used to measure the statistical performance of each method, with MCSEs presented in parentheses alongside the corresponding performance metrics.

In the full overlap scenarios, absolute bias is similarly low for MIM and the standard version of model-based standardization. In the limited overlap scenarios, the bias of both standardization methods has slightly higher magnitude. There seems to be a minimal increase in bias as the number of subjects in the index trial decreases. Bias is more marked in the scenarios with $N = 500$ (-0.019 and -0.013 for MIM and the standard approach, respectively, in the limited overlap setting, and -0.017 and -0.017, respectively, in the full overlap setting). This is likely due to the small-sample bias inherent in logistic regression [67].

As expected, precision is lost as the index trial sample size and the level of covariate overlap are reduced. With $\kappa = 0.5$, there exists a subpopulation within the target population that does not overlap with the index trial population. Therefore, inferences in a subsample of the target covariate dataset will rely on extrapolation of the conditional outcome model. With poorer covariate overlap, further extrapolation is required, thereby incurring a sharper loss of precision. Precision is very similar for both standardization methods, as ESEs are virtually equal in all simulation scenarios for both. Similarly, efficiency is virtually identical for both standardization methods. As per the ESE, MSE values also increase as the number of subjects in the index trial and the level of overlap decrease. Because bias is almost negligible across the simulation scenarios, efficiency is driven more by precision than by bias.

From a frequentist viewpoint, the empirical coverage rate should be equal to the nominal coverage rate to obtain appropriate type I error rates for null hypothesis testing. Namely, 95% interval estimates should include the true marginal log odds ratio 95% of the time. Theoretically, due to our use of 1,000 Monte Carlo simulations per scenario, the empirical coverage rate is statistically significantly different to the desired 0.95 if it is less than 0.9365 or more than 0.9635. For both standardization methods, empirical coverage rates only fall outside these boundaries, marginally — 0.934 for MIM and 0.935 for the standard approach – in the scenario with $N = 500$ and full overlap. This suggests that uncertainty quantification by the standardization methods is adequate.

Discussion

Despite measuring statistical performance in terms of frequentist finite-sample properties, MIM offers performance comparable to that of the standard version of model-based standardization. Both approaches provide appropriate inference with a correctly specified parametric conditional outcome model. The simulation study demonstrates proof-of-principle for the standardization methods, but only considers a simple best-case scenario with correct model specification and two continuous covariates. It does not investigate how robust the methods are to failures in assumptions.

Parametric outcome models impose strong functional form assumptions; for example, that effects are linear and additive on some transformation of the conditional outcome expectation. Such modeling assumptions may not be plausible where there are a large number of covariates and complex non-linear relationships between them. To provide some protection against model misspecification bias, one may consider using flexible data-adaptive estimators, e.g. non-parametric or machine learning techniques, for the conditional outcome model (the first-stage regression in MIM). While such approaches make weaker modeling assumptions, they may still be subject to larger-than-desirable bias in finite samples and are constrained by limited theoretical justification for valid statistical inference [68].

In practice, the use of MIM is appealing for several reasons. Firstly, as illustrated in Additional file 1, MIM can readily handle missingness in the patient-level data for the comparative index study. Missing outcomes, and potentially covariate and treatment values, could be imputed in each MCMC iteration of the synthesis stage, naturally accounting for the uncertainty in the missing data of the index study.

Secondly, the Bayesian first-stage regression model can incorporate both hard external evidence (e.g. the results of a meta-analysis) and soft external evidence (e.g. expert knowledge) to construct informative prior distributions for the model coefficients. When external data cannot be leveraged, "weakly informative" contextual information can be used to construct skeptical or regularization prior distributions. Through shrinkage, such priors can improve efficiency with respect to maximum-likelihood estimators in certain scenarios [69].

Thirdly, a Bayesian formulation for the first-stage regression offers additional flexibility to address other issues, such as measurement error in the patient-level data of the index trial [70]. Bayesian model averaging can be used to capture structural or model uncertainty [71]. When one is unsure about which baseline covariates are (conditional) effect measure modifiers, one can allow interactions to be "half in, half out" by specifying skeptical prior distributions for the candidate product term coefficients in Eq. 2 [72–74].[4]

[4] In the words of Simon and Freedman [74], this "encourages the quantification of prior belief about the size of interactions that may exist. Rather than forcing the investigator to adopt one of two extreme positions regarding interactions, it provides for the specification of intermediate positions."

In this article, we have used multiple imputation to perform model-based standardization over a target empirical covariate distribution, assumed to belong to participants that are external to the index study. In this scenario, one is reliant on correct specification of the outcome model (the first-stage regression in MIM) for unbiased estimation in the target. One is also reliant on covariates being consistently defined across data sources and on complete information on influential covariates being available both for the index study and for the target. In practice, this is a key challenge [75, 76], which could be addressed through the development of core patient characteristic sets that define clinically important covariates to be measured and reported among specific therapeutic areas [77].

In the absence of overlap between the covariate distributions in the index study and the external target, e.g. when the index study covariate distribution lies outside the target covariate distribution, one must consider the plausibility of the outcome model extrapolation. Sensitivity analyses using alternative model specifications may be warranted to explore the dependence of inferences on the selected adjustment model. Recently, several authors have proposed sensitivity analyses that are applicable where potential effect measure modifiers are measured only in the index trial but not in the target dataset [78, 79]. These techniques could be applied in conjunction with MIM.

While we have used multiple imputation to standardize over an external covariate distribution, it is also possible to standardize over the empirical covariate distribution of the index study, as illustrated in Additional file 1. This involves less stringent assumptions and avoids model-based extrapolation into an external data source. Such approach allows for the estimation of covariate-adjusted marginal treatment effects within individual comparative studies, adjusting for covariate imbalances between treatment arms.

A limitation of this article is the lack of a real case study demonstrating the application of the new methodology. While proof-of-principle for MIM has been provided through simulation studies, the method should be applied to a real example in order to influence applied practice. This is a key priority for future research.

Abbreviations

ADEMP	Aims, Data-generating mechanisms, Estimands, Methods, Performance measures
DAG	Directed acyclic graph
ESE	Empirical standard error
IPD	Individual patient data
HTA	Health technology assessment
MCMC	Markov chain Monte Carlo
MCSE	Monte Carlo standard error
MIM	Multiple imputation marginalization
MSE	Mean square error
RCT	Randomized controlled trial

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02157-x.

Additional file 1.

Acknowledgements
Not applicable.

Authors' contributions
A.R.A., A.H. and G.B. conceived the research idea, developed the methodology, performed the analyses, prepared the figures, and wrote and reviewed the manuscript.

Funding
AH is funded by Canada Research Chair in Statistical Trial Design; Natural Sciences and Engineering Research Council of Canada (award No. RGPIN-2021-03366).

Availability of data and materials
The files required to generate the data, run the simulations, and reproduce the results are available at http://github.com/remiroazocar/MIM.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Statistics and Data Insights, Bayer plc, 400 South Oak Way, Reading, UK. [2]Child Health Evaluative Sciences, The Hospital for Sick Children, 686 Bay Street, Toronto, Canada. [3]Dalla Lana School of Public Health, University of Toronto, 115 College Street, Toronto, Canada. [4]Department of Statistical Science, University College London, 1-19 Torrington Place, London, UK.

Received: 13 May 2023 Accepted: 19 January 2024
Published online: 10 February 2024

References

1. Remiro-Azócar A. Target estimands for population-adjusted indirect comparisons. Stat Med. 2022;41(28):5558–69.
2. Russek-Cohen E. Discussion of "target estimands for population-adjusted indirect comparisons" by Antonio Remiro-Azocar. Stat Med. 2022;41(28):5573–6.
3. Spieker AJ. Comments on the debate between marginal and conditional estimands. Stat Med. 2022;41(28):5589–91.
4. Senn S. Conditions for success and margins of error: estimation in clinical trials. Stat Med. 2022;41(28):5586–8.
5. Schiel A. Commentary on "Target estimands for population-adjusted indirect comparisons". Stat Med. 2022;41(28):5570–2.
6. Van Lancker K, Vo TT, Akacha M. Estimands in heath technology assessment: a causal inference perspective. Stat Med. 2022;41(28):5577–85.
7. Greenland S, Pearl J. Adjustments and their consequences-collapsibility analysis using graphical models. Int Stat Rev. 2011;79(3):401–26.
8. Greenland S, Morgenstern H. Confounding in health research. Annu Rev Public Health. 2001;22(1):189–212.
9. Kaufman JS. Marginalia: comparing adjusted effect measures. Epidemiology. 2010;21(4):490–3.
10. Whittemore AS. Collapsibility of multidimensional contingency tables. J R Stat Soc: Ser B (Methodol). 1978;40(3):328–40.

11. Greenland S, Pearl J, Robins JM. Confounding and collapsibility in causal inference. Stat Sci. 1999;14(1):29–46.
12. Huitfeldt A, Stensrud MJ, Suzuki E. On the collapsibility of measures of effect in the counterfactual causal framework. Emerg Themes Epidemiol. 2019;16:1–5.
13. Morris TP, Walker AS, Williamson EJ, White IR. Planning a method for covariate adjustment in individually-randomised trials: a practical guide. Trials. 2022;23(328).
14. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. Stat Med. 2013;32(16):2837–49.
15. Remiro-Azócar A, Heath A, Baio G. Methods for population adjustment with limited access to individual patient data: A review and simulation study. Res Synth Methods. 2021;12(6):750–75.
16. Remiro-Azócar A, Heath A, Baio G. Parametric G-computation for compatible indirect treatment comparisons with limited individual patient data. Res Synth Methods. 2022;13(6):716–44.
17. Remiro-Azócar A. Two-stage matching-adjusted indirect comparison. BMC Med Res Methodol. 2022;22(1):1–16.
18. Josey KP, Berkowitz SA, Ghosh D, Raghavan S. Transporting experimental results with entropy balancing. Stat Med. 2021;40(19):4310–26.
19. Phillippo DM, Dias S, Ades A, Belger M, Brnabic A, Schacht A, et al. Multilevel network meta-regression for population-adjusted treatment comparisons. J R Stat Soc Ser A (Stat Soc). 2020;183(3):1189–210.
20. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. Math Model. 1986;7(9–12):1393–512.
21. Zhang Z. Estimating a marginal causal odds ratio subject to confounding. Commun Stat-Theory Methods. 2008;38(3):309–21.
22. Moore KL, van der Laan MJ. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. Stat Med. 2009;28(1):39–64.
23. Austin PC. Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. J Clin Epidemiol. 2010;63(1):2–6.
24. Rosenblum M, Van Der Laan MJ. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. Int J Biostat. 2010;6(1).
25. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. Am J Epidemiol. 2011;173(7):731–8.
26. Wang A, Nianogo RA, Arah OA. G-computation of average treatment effects on the treated and the untreated. BMC Med Res Methodol. 2017;17(1):1–5.
27. Daniel R, Zhang J, Farewell D. Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. Biom J. 2021;63(3):528–57.
28. Campbell H, Park JE, Jansen JP, Cope S. Standardization allows for efficient unbiased estimation in observational studies and in indirect treatment comparisons: a comprehensive simulation study. 2023. arXiv preprint arXiv:230109661.
29. Vo TT, Porcher R, Chaimani A, Vansteelandt S. A novel approach for identifying and addressing case-mix heterogeneity in individual participant data meta-analysis. Res Synth Methods. 2019;10(4):582–96.
30. Rubin DB. Multiple imputation for nonresponse in surveys. vol. 81. New York: Wiley; 2004.
31. Westreich D, Edwards JK, Cole SR, Platt RW, Mumford SL, Schisterman EF. Imputation approaches for potential outcomes in causal inference. Int J Epidemiol. 2015;44(5):1731–7.
32. Remiro Azócar A. Population-Adjusted Indirect Treatment Comparisons with Limited Access to Patient-Level Data. London: UCL (University College London); 2022.
33. Remiro-Azócar A, Heath A, Baio G. Marginalization of regression-adjusted treatment effects in indirect comparisons with limited patient-level data. 2020. arXiv preprint arXiv:200805951.
34. Girman CJ, Ritchey ME, Zhou W, Dreyer NA. Considerations in characterizing real-world data relevance and quality for regulatory purposes: a commentary. Pharmacoepidemiol Drug Saf. 2019;28(4):439.
35. Weiss NS. Generalizing from the results of randomized studies of treatment: Can non-randomized studies be of help? Eur J Epidemiol. 2019;34(8):715–8.
36. Ramsey SD, Adamson BJ, Wang X, Bargo D, Baxi SS, Ghosh S, et al. Using electronic health record data to identify comparator populations for comparative effectiveness research. J Med Econ. 2020;23(12):1618–22.
37. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet. 2005;365(9453):82–93.
38. Rothwell PM. Commentary: External validity of results of randomized trials: disentangling a complex concept. Int J Epidemiol. 2010;39(1):94–6.
39. Greenhouse JB, Kaizar EE, Kelleher K, Seltman H, Gardner W. Generalizing from clinical trial data: a case study. The risk of suicidality among pediatric antidepressant users. Stat Med. 2008;27(11):1801–1813.
40. Happich M, Brnabic A, Faries D, Abrams K, Winfree KB, Girvan A, et al. Reweighting randomized controlled trial evidence to better reflect real life-a case study of the Innovative Medicines Initiative. Clin Pharmacol Ther. 2020;108(4):817–25.
41. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688.
42. Tackney MS, Morris T, White I, Leyrat C, Diaz-Ordaz K, Williamson E. A comparison of covariate adjustment approaches under model misspecification in individually randomized trials. Trials. 2023;24(1):1–18.
43. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res. 2011;46(3):399–424.
44. Meng XL. Multiple-imputation inferences with uncongenial sources of input. Stat Sci. 1994;9(4):538–58.
45. Gabrio A, Mason AJ, Baio G. A full Bayesian model to handle structural ones and missingness in economic evaluations from individual-level data. Stat Med. 2019;38(8):1399–420.
46. Rubin DB. Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In: Proceedings of the survey research methods section of the American Statistical Association. vol. 1. USA: American Statistical Association Alexandria; 1978. p. 20–34.
47. Vo T-T. A cautionary note on the use of G-computation in population adjustment. Res Synth Methods. 2023;14(3):338–41.
48. Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernan MA. Extending inferences from a randomized trial to a new target population. Stat Med. 2020;39(14):1999–2014.
49. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: a probabilistic programming language. J Stat Softw. 2017;76(1).
50. Raghunathan TE, Reiter JP, Rubin DB. Multiple imputation for statistical disclosure limitation. J Off Stat. 2003;19(1):1.
51. Rubin DB. Statistical disclosure limitation. J Off Stat. 1993;9(2):461–8.
52. Reiter JP. Satisfying disclosure restrictions with synthetic data sets. J Off Stat. 2002;18(4):531.
53. Reiter JP. Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. J R Stat Soc Ser A (Stat Soc). 2005;168(1):185–205.
54. Si Y, Reiter JP. A comparison of posterior simulation and inference by combining rules for multiple imputation. J Stat Theory Pract. 2011;5(2):335–47.
55. Reiter JP, Raghunathan TE. The multiple adaptations of multiple imputation. J Am Stat Assoc. 2007;102(480):1462–71.
56. Raab GM, Nowok B, Dibben C. Practical data synthesis for large samples. J Priv Confidentiality. 2016;7(3):67–97.
57. Bujkiewicz S, Achana F, Papanikos T, Riley R, Abrams K. Multivariate meta-analysis of summary data for combining treatment effects on correlated outcomes and evaluating surrogate endpoints. NICE DSU Tech Support Doc. 2019;20.
58. Reiter JP. Inference for partially synthetic, public use microdata sets. Surv Methodol. 2003;29(2):181–8.
59. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074–102.
60. Team RC, et al. R: a language and environment for statistical computing. 2013.
61. Phillippo DM, Dias S, Ades A, Welton NJ. Assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study. Stat Med. 2020;39(30):4885–911.
62. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. Stat Med. 2007;26(16):3078–94.

63. Austin PC, Stafford J. The performance of two data-generation processes for data with specified marginal treatment odds ratios. Commun Stat-Simul Comput®. 2008;37(6):1039–1051.
64. Remiro-Azócar A. Purely prognostic variables may modify marginal treatment effects for non-collapsible effect measures. 2022. arXiv preprint arXiv:221001757.
65. Goodrich B, Gabry J, Ali I, Brilleman S. rstanarm: Bayesian applied regression modeling via Stan. R Pack Version. 2020;2(1).
66. Team SD. RStan: the R interface to Stan. R Packag Version. 2020;2(21.2).
67. Nemes S, Jonasson JM, Genell A, Steineck G. Bias in odds ratios by logistic regression modelling and sample size. BMC Med Res Methodol. 2009;9:1–5.
68. Naimi AI, Mishler AE, Kennedy EH. Challenges in obtaining valid causal effect estimates with machine learning algorithms. Am J Epidemiol. 2023;192(9):1536–44.
69. Keil AP, Daza EJ, Engel SM, Buckley JP, Edwards JK. A Bayesian approach to the g-formula. Stat Methods Med Res. 2018;27(10):3183–204.
70. Keil AP, Daniels JL, Hertz-Picciotto I. Autism spectrum disorder, flea and tick medication, and adjustments for exposure misclassification: the CHARGE (CHildhood Autism Risks from Genetics and Environment) case-control study. Environ Health. 2014;13(1):1–10.
71. Madigan D, Raftery AE. Model selection and accounting for model uncertainty in graphical models using Occam's window. J Am Stat Assoc. 1994;89(428):1535–46.
72. Dixon DO, Simon R. Bayesian subset analysis. Biometrics. 1991;47(3):871–81.
73. Spiegelhalter DJ, Freedman LS, Parmar MK. Bayesian approaches to randomized trials. J R Stat Soc Ser A (Stat Soc). 1994;157(3):357–87.
74. Simon R, Freedman LS. Bayesian design and analysis of two x two factorial clinical trials. Biometrics. 1997;53(2):456–64.
75. Stuart EA, Bradshaw CP, Leaf PJ. Assessing the generalizability of randomized trial results to target populations. Prev Sci. 2015;16:475–85.
76. Stuart EA, Rhodes A. Generalizing treatment effect estimates from sample to population: A case study in the difficulties of finding sufficient data. Eval Rev. 2017;41(4):357–88.
77. Vuong ML, Tu PHT, Duong KL, Vo TT. Development of minimum reporting sets of patient characteristics in epidemiological research: a methodological systematic review. Res Methods Med Health Sci. 2023;26320843231191777.
78. Nguyen TQ, Ebnesajjad C, Cole SR, Stuart EA. Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. Ann Appl Stat. 2017;11(1):225–47.
79. Dahabreh IJ, Hernán MA. Extending inferences from a randomized trial to a target population. Eur J Epidemiol. 2019;34:719–22.

Publisher's Note