Biometrics, 2025, 81(1), ujae167
https://doi.org/10.1093/biomtc/ujae167
Biometric Practice

Estimating hypothetical estimands with causal inference and missing data estimators in a diabetes trial case study

Camila Olarte Parra [1,*], Rhian M. Daniel [2], David Wright[3], Jonathan W. Bartlett [4]

[1]Unit of Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Nobels väg 13, Solna, Stockholm 171 65, Sweden, [2]Division of Population Medicine, Cardiff University, Cardiff CF14 4YS, United Kingdom, [3]Respiratory and Immunology Biometrics and Statistical Innovation, BioPharmaceuticals R&D, AstraZeneca, Cambridge CB2 0AA, United Kingdom, [4]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London WC1E 7HT, United Kingdom

[*]Corresponding author: Camila Olarte Parra, Unit of Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Nobels väg 13, Solna, Stockholm 171 65, Sweden (camila.olarte.parra@ki.se).

ABSTRACT

The ICH E9 addendum on estimands in clinical trials provides a framework for precisely defining the treatment effect that is to be estimated, but says little about estimation methods. Here, we report analyses of a clinical trial in type 2 diabetes, targeting the effects of randomized treatment, handling rescue treatment and discontinuation of randomized treatment using the so-called hypothetical strategy. We show how this can be estimated using mixed models for repeated measures, multiple imputation, inverse probability of treatment weighting, G-formula, and G-estimation. We describe their assumptions and practical details of their implementation using packages in R. We report the results of these analyses, broadly finding similar estimates and standard errors across the estimators. We discuss various considerations relevant when choosing an estimation approach, including computational time, how to handle missing data, whether to include post intercurrent event data in the analysis, whether and how to adjust for additional time-varying confounders, and whether and how to model different types of intercurrent event data separately.

KEYWORDS: causal inference; E9 addendum; hypothetical estimand; intercurrent events; missing data.

1 INTRODUCTION

Following the ICH E9 addendum, defining the treatment effect of a clinical trial, also known as the "estimand", includes identifying intercurrent events (ICEs) and strategies to deal with them (ICH, 2019). Examples of ICE include treatment discontinuation, rescue medication use, death prior to measuring the outcome, or any event that occurs after treatment initiation that either affects the interpretation or the existence of the outcome.

In diabetes trials, rescue medication for adequate glucose control should be available for ethical reasons because of the deleterious effect of elevated glucose levels. One option is to target the treatment effect in a way that includes any effect that the addition of rescue medication may have on the outcome. According to the ICH E9 addendum, this would correspond to using a *treatment policy* strategy to deal with this ICE. This strategy, however, leads to an estimand that may mask (or, less commonly, exaggerate) the effect of the study drug itself whenever there is differential use of rescue medication between treatment arms (Holzhauer et al., 2015). In particular, if there is a higher rescue use in the control compared to the active arm (and if the rescue is more effective than the control arm medication alone), the treatment policy estimand may understate the pharmacological benefits of the active treatment. Estimating the treatment effect in the (hypothetical) absence of rescue medication use can then be of interest for certain stakeholders. In this case, the use of rescue medication would be handled following a so-called *hypothetical* strategy, targeting what would have been observed in the trial had rescue medication not been made available to patients (even if contrary to the fact). It is important to note, however, that this is just one possible hypothetical estimand that one could contemplate (Lipkovich et al., 2020).

In Section 2, a trial in type 2 diabetes patients is described as a motivating example. The choice of statistical analysis in the published analysis suggests that the primary estimand of interest would have used a hypothetical strategy to deal with rescue medication and treatment discontinuation. This estimand is the main focus of this paper and our aim is to describe and illustrate how different estimators can be applied in a real life scenario. Missing data methods are typically used to estimate such estimands, because the hypothetical outcome values that would have ensued in the absence of the ICE are incomplete. These include mixed-model repeated measures (MMRMs) and multiple imputation (MI). Causal inference estimators, like G-formula, inverse probability of treatment weighting (IPTW) and G-estimation, have thus far been rarely used in clinical trials, presumably because they were mostly developed for observational rather than randomized studies. In earlier work, we showed how causal inference estimators can be used to estimate hypothetical estimands, with the potential for improved statistical efficiency over

Received: September 1, 2023; Revised: October 19, 2024; Accepted: January 3, 2025

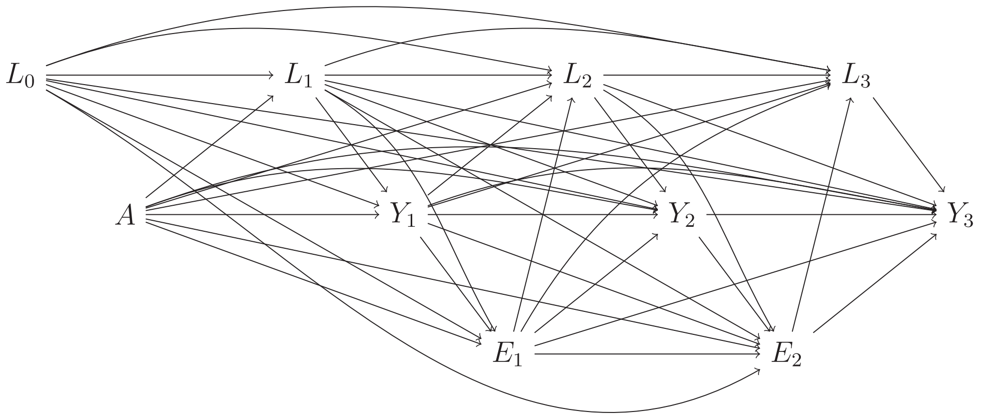

FIGURE 1 Directed Acyclic Graph relating randomized treatment (A), baseline (L_0) and time-varying covariates (L_k), occurrence of the ICE at each visit (E_k) and repeated measurements of the outcome (Y_k).

estimates obtained using missing data methods (Olarte Parra et al., 2022). Here, we demonstrate feasible ways to implement the different estimators using existing statistical packages and discuss how to tackle challenges encountered in real life settings including missing data. We also describe and contrast the statistical assumptions and properties of each estimator, and based on this give recommendations for trial statisticians faced with the decision of which to use.

2 MOTIVATING EXAMPLE

We analyzed data from a trial where type 2 diabetes patients on metformin monotherapy who had inadequate glycemic control were randomized at baseline to additionally receive dapagliflozin, dapagliflozin, and saxagliptin, or glimepiride (Müller-Wieland et al., 2018). Their HbA1c and fasting plasma glucose were measured at baseline and then periodically to assess their response. During the first 3 months, visits occurred every 2 weeks and then every 12 weeks up to 52 weeks. The main outcome of interest (Y) was change in HbA1c from baseline to the final visit after 52 weeks of follow-up. Insulin was indicated as rescue medication for patients with inadequate glycemic control. Up to visit 9, rescue was considered if patients' fasting plasma glucose (FPG) exceeded a visit specific threshold, while at visits 10–12, rescue was considered if their HbA1c exceeded a specified threshold. If patients exceeded the relevant threshold they then had an extra visit and their FPG or HbA1c value was measured again. Rescue was then initiated if this value exceeded the threshold. Once rescue medication was started, patients continued taking it for the remainder of the study. There were no deaths observed during the study period.

For simplicity, we will focus on the comparison of dapagliflozin and saxagliptin to glimepiride. We chose this comparison to illustrate the potential effect of an imbalance in use of rescue medication, given that in the glimeperide arm 12.3% received rescue medication while in the dapagliflozin and saxagliptin arm only 6.2% received it. Access to this trial data were approved by the sponsor AstraZeneca and requested via the platform Vivli (Vivli Data Request: 6764).

3 ESTIMAND AND IDENTIFIABILITY ASSUMPTIONS

To precisely define our estimand of interest, we will introduce some notation. Let A denote randomized treatment and E_k denote a binary indicator whether by visit k the patient had had an ICE (of either type). Patients who started rescue or discontinued treatment, remained on rescue or without study treatment for the rest of the follow up. The overbar denotes the history of a variable up to and including visit k (eg $\overline{E}_k$) or throughout the entire follow-up (up to and including visit 9), for example $\overline{E}$. We let Y_0 denote HbA1c at baseline. $Y_k^{a,\bar{e}_{k-1}}$ denotes the *potential outcome* for change in HbA1c from baseline that would be measured at visit k if we were to set treatment $A = a$ and ICE occurrence $\overline{E}_{k-1} = \bar{e}_{k-1}$. We emphasize that such potential outcomes are assumed to be well-defined for all patients, regardless of their actual treatment assignment A and actual ICE occurrence up to visit k, $\overline{E}_k$ (Rubin, 1974). In particular, implicit in the notation $Y_k^{a,\bar{e}_{k-1}}$ is that a given patient's potential outcome does not depend on the treatment (and in our setting, ICE values) assigned to any other patients (no-interference) and that there do not exist multiple (sub)versions of treatment that might give rise to different outcomes (no-versions-of-treatment) (Rubin, 1980).

Our estimand of interest is then

$$\mathbf{E}\left(Y_{10}^{a=1,\bar{e}=\bar{0}} - Y_{10}^{a=0,\bar{e}=\bar{0}}\right). \tag{1}$$

In words, this is the effect (as a mean difference) on the change in HbA1c measured at the final (10th) visit (Y_{10}) of dapagliflozin and saxagliptin $(A = 1)$ compared to glimeperide $(A = 0)$ as add-on medications, if rescue medication had not been made available and patients had continued taking their assigned treatment during the full follow-up $(\overline{E} = \overline{0})$. The estimand is an example of a controlled direct effect—it is the direct effect of randomized treatment on outcome not mediated via the treatment's effect on the ICE, and where the mediator (the ICE) is controlled at a specific level (here zero) (Hernán and Robins, 2024).

Figure 1 summarizes the assumed causal structure between randomized treatment A, repeated measurements of outcome Y and the occurence of the ICE in a simplified setting with 3 visits after baseline. L_0 denotes variables measured at baseline and L_k

TABLE 1 Summary of the different estimators and their implementation in R.

Method	R package	Data format	Prior MI to handle missing data	Post-ICE values	Standard error
MMRM	`mmrm`	Long	No	No	Likelihood-based
MI	`mice`	Wide	Yes	No	Rubin's rules
IPTW	`ipw`	Long	Yes	No	Bootstrap
G-formula	`gfoRmula`	Long	No	Yes	Bootstrap
G-formula via MI	`mice`	Wide	Yes	Yes	Raghunathan's formula
G-estimation	None	Wide	Yes	Yes	Bootstrap

variables measured at each follow-up visit k. The choice of which variables to include in L_0 and L_k should be made in order to render the following identifiability assumptions plausible:

(1) Consistency: for those randomized to arm a who do not experience an ICE, their observed outcome is equal to their potential outcome $Y_{10} = Y_{10}^{a,\bar{e}=\bar{0}}$.

(2) Sequential exchangeability: together with A and $\overline{E}_{k-1}$, $\overline{L}_k$ and $\overline{Y}_k$ are sufficient to control for confounding between each E_k and Y_{10}. Formally, $Y_{10}^{a,\bar{e}=\bar{0}} \perp E_k | A = a, \overline{L}_k, \overline{Y}_k, \overline{E}_{k-1} = \bar{0}$ for all k and for $a = 0, 1$.

(3) Positivity: if a participant is ICE-free up to visit $k-1$, there is a positive probability of again *not* having the ICE at visit k, conditional on any possible combination of covariate (and past outcome) history and treatment arm: $P(E_k = 0 | A = a, \overline{L}_k = \bar{l}_k, \overline{Y}_k = \bar{y}_k, \overline{E}_{k-1} = \bar{0}) > 0$ for all k, for $a = 0, 1$ and for any $(\bar{l}_k, \bar{y}_k)$ such that the conditional covariate (and past outcome) density given treatment arm and "no ICE" is bounded away from zero at $(\bar{l}_k, \bar{y}_k)$. Note that assumptions such as $P(E_k = 1 | A = a, \overline{L}_k = \bar{l}_k, \overline{Y}_k = \bar{y}_k, \overline{E}_{k-1} = \bar{0}) > 0$ are not required because we are only interested in ICE-free potential outcomes.

We note that the consistency, no-interference and no-versions-of-treatment assumptions together make up what Rubin termed the Stable Unit Treatment Value Assumption (Rubin, 1980; VanderWeele and Hernan, 2013). The above identifiability conditions allow the estimand to be expressed as a function of the observed data. However, unless all variables are discrete, statistical inference requires use of additional statistical modelling assumptions (Hernán and Robins, 2024), which we describe further in Section 4 for each estimator.

Ideally, when defining such an estimand for a future trial, one would consult with experts in the corresponding therapeutic area at protocol stage to ensure the collection of variables such that the identifiability conditions (specifically exchangeability and positivity) will be plausibly satisfied. In our case of retrospective analysis of an already conducted trial, we instead necessarily chose the variables to go into L_0 and L_k based on what was collected and thus available from the original trial, our subject matter knowledge, and the trial's protocol.

We chose the baseline covariates L_0 to consist of age, sex, body mass index (BMI), systolic blood pressure, duration of diabetes and C-peptide (indicator of the production of insulin), which were presented in Table 1 in the original trial publication. In the version of the dataset to which we had access, age was grouped into 5-year categories, except for the first category that included patients from 18 to 30 years ($n = 10$). We used the mid-point of each category to create a variable that was treated as a continuous variable in these analyses. When doing the primary analysis of a trial, this would not be necessary as the actual age would be available; this approximation is only to facilitate the analysis here. BMI was available only as 3 categories (normal weight, overweight and obesity) and thus was included as categorical.

The time-varying covariates L_k were chosen to be FPG and kidney function (estimated glomerular filtration rate, eGFR). As we explained in Section 2, rescue medication was indicated according to FPG and HbA1c. We chose to include only their scheduled measurements and not the additional measurements taken in those patients whose planned measure exceeded the threshold at the scheduled visit. This was decided to avoid possible violations of the positivity assumption, and because omission of these measurements would only violate the exchangeability assumption if they exerted a large direct effect (ie, not through the effect on subsequent rescue initiation) on the outcome, which seems unlikely. We also accounted for kidney function via inclusion of eGFR because impairment can lead to the discontinuation of these medications. As with other baseline characteristics, including repeated measurements of FPG and eGFR makes the exchangeability assumption more plausible.

4 METHODS

In this section, we describe the different estimators used to estimate the hypothetical estimand defined in Section 3. For each, we describe and contrast their statistical assumptions and how our variable and modeling choices were made to increase the chances these assumptions were satisfied. We start by replicating the original analysis with MMRM and then provide alternatives that include MI, IPTW, G-formula, and G-estimation. We discuss variations of these approaches that include other relevant variables besides the ones included in the original analysis. The different methods and implementations are numbered sequentially as they are being described to link them with their corresponding result in Section 5. The more detailed step by step implementation of each method and the corresponding software and code used to implement them can be found in the online Supplementary Material.

4.1 Methods using only values before the ICE

4.1.1 Mixed-model repeated measures

The original paper publishing the trial results was based on MMRM, which treats the unobserved no-ICE outcomes in those patients who experienced an ICE as missing data. In general, in line with recommendations of the ICH E9 addendum (ICH, 2019), we should use in the analysis information from all randomized patients. However, in the original trial, results were based on the "full analysis set", defined as the set of patients who received at least one dose of the treatment, had a baseline value of HbA1c, and at least one follow-up measurement. We restricted our analyses to this subset so that any differences in results from those in the original paper are not due to a different choice of patients used in the analysis. This same set of participants was used for all the analyses we conducted. Only measurements of the outcome (change in HbA1c) before rescue treatment or discontinuation were included. This means that for individuals who had either ICE, we are setting Y to missing at visits after the ICE occurred, even if it was actually observed. The MMRM model (*Method 1*) specifies fixed effects of treatment, visit, and their interaction, baseline HbA1C and its interaction with visit, and an unstructured residual error covariance matrix. The model thus assumed

$$Y_k^{a,\bar{e}=0} = \alpha_{0,k} + \alpha_{A,k}a + \alpha_{Y_0,k}Y_0 + \epsilon_k \qquad (2)$$

for $k = 1, \dots, K$, with $(\epsilon_1, \dots, \epsilon_K)^T \sim N(0, \Sigma)$ and where Σ denotes an arbitrary (unstructured) positive definite covariance matrix. MMRMs maximize the observed data likelihood, and provide valid inferences when the full data (here the hypothetical no-ICE outcomes) model is correctly specified and the unobserved no-ICE potential outcome data satisfy the missing at random (MAR) assumption. This MAR assumption, which in the case of monotone missingness is the same as the sequential exchangeability assumption (Olarte Parra et al., 2022), is violated by the presence of the variables L_k which affect HbA1C and the ICE but which are not used in the MMRM model. However, the extent to which it is violated may be mitigated by the high correlation between FPG and HbA1c (Holzhauer et al., 2015). In principle, one could fit an extended MMRM model which includes the FPG measurements at each visit as part of the outcome vector. However, this approach is not suitable whenever the time-varying confounders are such that the resulting multivariate normal assumption would not be plausible (eg, if one were binary). The MI approach we describe next offers a more flexible and convenient approach to incorporate variables like FPG which affect the outcome of interest and also ICE occurrence.

4.1.2 Multiple imputation

Using the same data as in the MMRM analysis and assuming MAR, the no-ICE potential outcomes were imputed using MI. We used the MI by chained equations (or fully conditional specification) approach to impute the missing no-ICE values (*Method* 2). The base imputation model for $Y_k^{\bar{e}=0}$ assumed

$$Y_k^{a,\bar{e}=0} = \beta_{0,k} + \beta_{A,k}a + \beta_{Y_0,k}Y_0 + \beta_{Y_{-k},k}^T Y_{-k}^{a,\bar{e}=0} + \epsilon_k, \quad (3)$$

where $Y_{-k}^{a,\bar{e}=0} = (Y_1^{a,\bar{e}=0}, \dots, Y_{k-1}^{a,\bar{e}=0}, Y_{k+1}^{a,\bar{e}=0}, \dots, Y_K^{a,\bar{e}=0})$ denotes the vector of changes in HbA1c (under no-ICE) from baseline except for visit k and $\epsilon_k \sim N(0, \sigma_k^2)$. This is equivalent to imputation from the joint multivariate model assumed in the MMRM model (Hughes et al., 2014). We do not expect these results to be numerically identical to MMRM for reasons discussed by Wang and Robins (1998) (eg, taking only a finite number of imputations); however, for a large sample size and a large number of imputations, the differences are expected to be very small. An advantage of MI compared to MMRM is that it allows the inclusion of a different set of variables in the imputation and analysis models. By including additional variables in the imputation model, the MAR assumption can be rendered more plausible. Thus, we also implemented other versions of MI that included the baseline (*Method* 3) and time-varying variables (*Method 4*) listed in Section 3 in the imputation models. As such, FPG and eGFR were imputed using normal imputation models analogous to Equation (3), and they served as covariates in the imputation models for HbA1C.

4.1.3 Inverse probability of treatment weighting

MMRM and MI rely on models for the outcomes (and time-varying confounders in the case of MI). For MI, correctly specifying all of these models may be particularly demanding with multiple time-varying confounders. An alternative approach which avoids this requirement is to instead model the ICE mechanism, which here corresponds to the missingness (in no-ICE outcomes) mechanism, and then use inverse probability of treatment/missingness weighting (IPTW, *Method* 5). For this setting, the time-varying "treatment" corresponds to the the occurrence of the ICE. The weights were estimated based on a pooled logistic regression model assuming

$$\begin{aligned} P(E_k = 0 | A, Y_0, \bar{Y}_k, L_0, \bar{L}_k, \bar{E}_{k-1} = 0) &= \text{expit}(\gamma_0 + \gamma_A A + \gamma_{Y_0} Y_0 + \gamma_Y Y_k \\ &\quad + \gamma_{Y_{-1}} Y_{k-1} + \gamma_{G_{-1}} G_{k-1}), \end{aligned} \qquad (4)$$

where G_{k-1} was the average HbA1c up to and including visit $k - 1$. For those patients who did not experience an ICE through the final follow-up, their weight was calculated as $V = \prod_{k=1}^{K} \frac{1}{P(E_k=0|A,Y_0,\bar{Y}_k,L_0,\bar{L}_k,\bar{E}_{k-1}=0)}$. Sometimes stabilized weights are used in IPTW for estimating parameters of marginal structural models, but here, where there is no such model, their use would make no difference (the numerator term would be identical among all patients who are ICE free through to the final follow-up visit). We chose to include HbA1c at the same visit Y_k, the previous one (Y_{k-1}) and the average of the earlier ones on the basis that this should constitute a parsimonious representation of how past HbA1c may have affected the occurrence of ICE. Since the trial was double-blind, in principle, we could have omitted the randomized treatment A from the model. We nonetheless included it since it is predictive of outcome if there is a treatment effect (Brookhart et al., 2006). Below, we explain further variations where additional covariates were included in the logistic model to increase the plausibility that sequential exchangeability was satisfied.

The weights for IPTW were estimated using the `ipwtm` function of the `ipw` R package (van der Wal and Geskus, 2011). An important issue to highlight is that the `ipw` package does not allow missing values. To overcome this, we imputed the missing values before applying IPTW. As with MI, we have the flexibility to include additional covariates in the model for the weights (Equation [4]) in order to make the sequential exchangeability assumption more plausible. Thus, we additionally conducted IPTW also including baseline covariates (*Method 6*) and time-varying variables without interactions (*Method 7*).

It is worth highlighting that in MMRM or MI, which do not use post-ICE data, there is no need and indeed one cannot distinguish between the types of ICE which occur in the models. This is because, by definition, there is no information in the observed data about how the type of ICE experienced by a patient might predict their unobserved no-ICE outcome. In contrast, for IPTW, as the ways covariates predict the occurrence of the 2 different ICE types may differ, an approach that distinguishes between the different ICE may be preferable. We could consider having separate logistic models for each ICE and then using the product of the probabilities of having each ICE to construct the weights. However, this would imply that the events are independent. Alternatively, we can construct the ICE variable as a factor variable to indicate whether no ICE occurred ($E_k = 0$), only rescue ($E_k = 1$), only discontinuation ($E_k = 2$) or both occurred ($E_k = 3$) and use this to estimate the weights by specifying a multinomial logistic regression model in the argument family of the `ipwtm` function with all baseline and time-varying variables as covariates (*Method 8*). Since the multinomial model only includes contributions up until a patient has non-zero ICE status and there were no patients who in the same visit were both rescued and discontinued, in fact only the probabilities of $E_k = 1$ and $E_k = 2$ versus $E_k = 0$ were modeled.

For this and the following methods that required the bootstrap, we chose to draw 100 bootstrap samples, unless otherwise specified. Often one would use a larger number of bootstrap samples to minimize the Monte-Carlo error in the bootstrap estimate of variance. We chose a relatively small value here because the bootstrap variance estimates are themselves averaged across the 100 imputed datasets to calculate the average within-imputation variance.

Estimating the effect on each imputed dataset and bootstrapping within each imputed dataset to obtain a within-imputation variance estimate is computationally faster than bootstrapping the observed dataset with missing values and then imputing in each bootstrap sample (Schomaker and Heumann, 2018). Moreover, Rubin's variance estimator using bootstrapping to estimate the within-imputation variance has previously been shown to work well when combining MI with inverse probability weighting (Leyrat et al., 2019).

4.2 Methods exploiting post-ICE values

An advantage of not using outcome values after the ICE occurs is that it avoids the need to model the effect of the ICE on the outcome. Nonetheless, to an increasing extent, trials continue to collect information on patients after experiencing ICEs. We now describe methods that can exploit such information, potentially increasing statistical efficiency, but at the expense of having to make additional modeling assumptions.

4.2.1 *Parametric G-formula*

The G-formula (*Method 9*) by default makes use of measurements taken after the ICE occurs, in contrast to the approaches described previously. G-formula fits models for the time-varying confounders L_k, outcomes Y_k, and the final outcome Y_{10}. It then simulates potential outcomes based on these and the treatment sequence of interest, which for the hypothetical estimand corresponds to setting the ICE to 0 throughout. Our first G-formula implementation did not use the time-varying confounders L_k, while for the outcomes Y_k, it assumed a pooled model of the form

$$\begin{aligned} Y_k = \delta_0 + \delta_A A + \delta_{E_{-1}} E_{k-1} + \delta_W W_k + \delta_{Y_0} Y_0 \\ + \delta_{Y_{-1}} Y_{k-1} + \delta_{G_{-1}} G_{k-1} + \epsilon_k, \end{aligned} \tag{5}$$

where $\epsilon_k \sim N(0, \sigma^2)$ and W_k denotes the number of weeks since the patient experienced the ICE, or 0 if no ICE had occurred by visit k or it had just occurred at visit k. We chose this, as opposed to the number of visits since the ICE first occurred, because the time interval between visits varied, as explained in Section 2. Our model specification thus allowed for an effect of having earlier had an ICE, and this effect was allowed to depend on how long since the ICE occurred. The terms Y_{k-1} and G_{k-1} were included as a parsimonious summary of how past HbA1C was assumed to affect current HbA1c. Relaxing this assumption and allowing independent effects of all past HbA1c values is more readily implemented using the G-formula via MI method we describe in Section 4.2.2.

To render the sequential exchangeability assumption being made more plausible, we also ran G-formula (*Method 10*) including additional covariates in the HbA1c model (step 2) and including FPG and GFR as time-varying confounders. Besides randomized treatment, the ICE indicator, the W_k ICE variable, current HbA1c, lagged HbA1c, and lagged average HbA1c, the models for FPG, GFR, and HbA1c included the same baseline and time-varying covariates for the previous methods listed in Section 3. We also included a lagged value and lagged average value of FPG and GFR in all the models. The final model using the simulated data (step 3) was not modified that is, it only included randomized treatment and baseline HbA1c as covariates with simulated change in HbA1c at the last visit as response variable. These G-formula implementations assume that the effect of an ICE on subsequent outcomes is the same irrespective of which type of ICE had occurred, which is likely false in reality. To accommodate this, we also included an additional implementation where the ICE was a categorical variable indicating that the ICE had not occurred ($E_k = 0$), only rescue ($E_k = 1$), only discontinuation ($E_k = 2$) or both had occurred ($E_k = 3$) (*Method 11*).

In contrast to the `ipw` package, the `gfoRmula` package can be used with datasets with missing values. A single model is fitted for each time-varying confounder to the pooled long-form data. Any rows (measurements of a patient at a given visit) in which a missing value occurs in the response or covariates in these models are ignored by default by R's regression model fitting functions. The resulting "complete case" fits yield valid estimates provided missingness is independent of the response variable, given the covariates. In general, this assumption does not coincide with an MAR assumption, and indeed it may often be deemed more plausible than MAR (White and Carlin, 2010).

4.2.2 *G-formula via MI*

An alternative approach to handling missing data when using G-formula is to use MI to first impute missing data, then use G-formula on each imputed dataset, pooling results using Rubin's rules. While this is possible, it is highly computationally intensive, partly because of the use of bootstrapping to obtain within-imputation variance estimates. To avoid this high computational burden, we also implemented G-formula by using synthetic data MI methods (*Method 12*), as proposed by Bartlett et al. (2023). This involves using Bayesian MI methods to both impute missing data and simulate the potential outcomes of interest. Because we used existing MI software to do this, which imputes data in the wide form, our implementation assumed when imputing no-ICE potential outcomes

$$Y_k = \lambda_{0,k} + \lambda_{A,k}A + \lambda^T_{\bar{E}_{k-1},k}\bar{E}_{k-1} + \lambda_{Y_0,k}Y_0 + \lambda^T_{\bar{Y}_{k-1},k}\bar{Y}_{k-1} + \epsilon_k, \tag{6}$$

where $\epsilon_k \sim N(0, \sigma_k^2)$. Thus, this G-formula implementation fitted separate models for HbA1c at each visit, relaxing the assumptions of common effects across visits made in Equation (5) with the gfoRmula package, and also allowing independent effects of all past HbA1c values.

Analogous to G-formula, to render the sequential exchangeability assumption more plausible, we implemented a version of G-formula via MI (*Method 13*) that included the baseline and time-varying covariates listed for the previous methods in both imputation models and another with all these covariates but categorical indicators of the ICE type (*Method 14*).

4.2.3 *G-estimation*

G-estimation (*Method 15*) is an alternative approach that has recently been used for estimating hypothetical estimands (Lasch and Guizzaro, 2022; Lasch et al., 2022). In this approach, outcomes are sequentially adjusted to remove the effects of the mediator, which under certain assumptions permits estimation of controlled direct effects, of which as noted earlier, the hypothetical estimand is an example (Loh et al., 2020). The approach is based on assuming a so-called structural nested mean model (SNMM), which specifies the effect of the mediator (here the ICE) being set to 1 at a given visit on the final outcome, setting the mediator to 0 at all subsequent visits (Vansteelandt and Sjolander, 2016). To define the SNMM we use, let M_k denote the binary indicator that the ICE occurs at (rather than by) visit k, and let $Y_{10}^{a,\overline{m}}$ denote the final outcome setting randomized treatment to a and $\overline{M}$ to $\overline{m}$. Let c_k denote a length 9 vector whose entries are 0 except the kth, which is 1. Then, our SNMM assumes that for $k = 1, \ldots, 9$

$$E\left[Y_{10}^{a,\overline{m}=c_k} - Y_{10}^{a,\overline{m}=0} | A = a, \overline{M}_{k-1} = \bar{0}, Y_0, \overline{Y}_{k-1}, \overline{L}_k\right] = \psi_k. \tag{7}$$

The parameter ψ_k captures the effect of having the ICE at visit k compared to never having the ICE, among those ICE free before visit k. It moreover assumes that this effect does not vary with the earlier values of the time-varying confounders and outcome. Under the assumed SNMM and the identification assumptions described in Section 3, it can be shown that $E(Y_{10}^{a,\overline{m}=0}) = E(Y_{10} - \sum_{k=1}^{k=9} \psi_k M_k | A = a)$, from which the hypothetical estimand can be estimated (Loh et al., 2020). The term $Y_{10} - \sum_{k=1}^{k=9} \psi_k M_k$ corresponds to an individual's outcome with the effects of the ICE removed, if they experienced the ICE. To estimate the parameters ψ_k, a series of regression models are fitted to the successively adjusted outcomes. Thus, whereas G-formula requires models for the time-varying confounders L_k, G-estimation of SNMM requires models for these mediator adjusted outcomes. As such, G-estimation of SNMM is arguably less demanding from a model specification perspective, particularly when, as is typically the case, L_k consists of multiple variables.

We also implemented a version of G-estimation where we included additional baseline and time-varying covariates (*Method 16*) and one with all the covariates and the categorical version of the ICE (*Method 17*).

Table 1 shows a summary of the different estimators with the corresponding R package used, the data format required, how missing data were handled, whether it included post-ICE values and how the corresponding SEs were estimated.

5 RESULTS

Table 2 summarizes the characteristics of the patients randomized to each treatment arm of the trial. There are fewer patients than in the original trial publication because some of them withdrew consent ($n = 33$). There were very few missing baseline values, with many variables having complete information and the rest having less than 0.5% missing per variable. It is worth noting that most of the missing values occurred in the dapagliflozin arm, that was not included in our analysis. Compared to the dapagliflozin + saxagliptin arm, there were more treatment discontinuations ($n = 14$, 4.6% vs. $n = 7$, 2.3%) and use of rescue medication ($n = 37$, 12.3% vs. $n = 19$, 6.2%) than in the glimeperide arm.

Table 3 summarizes the numbers of missing outcomes per visit in each arm. At each visit, there were more missing values in the glimeperide arm than in the dapagliflozin and saxagliptin arm. Treating post-ICE values as missing increases missing outcome values from 0.1%–5 to 3%–20% per visit. In visit 9, which has the higher number of missing values, the missingness increases from 5% to 15% when outcome values after the ICE are deleted.

In Section 2, we noted that rescue medication at visits 1–8 was indicated according to the values of two measurements of FPG, only the first of which is used in our analyses. The omission of the second measurement is expected to avoid positivity violations, although near-violations are still a concern. The figure in the Supplementary Material shows a plot of the (first) FPG values per visit with the colour indicating whether rescue was initiated either at that visit, or an earlier visit. The threshold (which changed between visits 6 and 7) used for the FPG measurement is indicated by the black line, and the distribution of

TABLE 2 Patient characteristics.

Variable	**Dapagliflozin 10mg (N=299)**	**Glimepiride 1mg/2mg/4mg (N=302)**	**Dapagliflozin 10mg and Saxagliptin 5mg (N=305)**	**Overall (N=906)**
Age (years), mean (SD)	56.9 (9.59)	58.2 (8.43)	58.8 (7.98)	58.0 (8.71)
Sex, *n* (%)				
Women	108 (36.1%)	98 (32.5%)	119 (39.0%)	325 (35.9%)
Men	191 (63.9%)	204 (67.5%)	186 (61.0%)	581 (64.1%)
Baseline body mass index (kg/m ^{2})				
$19 < x \leq 25$	9 (3.0%)	9 (3.0%)	20 (6.6%)	38 (4.2%)
$25 < x \leq 30$	76 (25.4%)	83 (27.5%)	87 (28.5%)	246 (27.2%)
$30 < x \leq 80$	212 (70.9%)	210 (69.5%)	198 (64.9%)	620 (68.4%)
Missing, n (%)	2 (0.7%)	0 (0%)	0 (0%)	2 (0.2%)
Baseline systolic blood pressure (mmHg), mean (SD)	138 (14.4)	139 (13.0)	139 (14.0)	139 (13.8)
Missing, *n* (%)	1 (0.3%)	0 (0%)	0 (0%)	1 (0.1%)
Baseline waist circumference (cm), mean (SD)	111 (14.0)	112 (13.2)	109 (12.4)	111 (13.2)
Missing, *n* (%)	2 (0.7%)	3 (1.0%)	2 (0.7%)	7 (0.8%)
Baseline hip circumference (cm), mean (SD)	113 (12.5)	112 (11.9)	111 (11.4)	112 (11.9)
Missing, *n* (%)	4 (1.3%)	2 (0.7%)	3 (1.0%)	9 (1.0%)
Years since first diagnose, mean (SD)	6.88 (5.24)	6.73 (5.14)	7.39 (5.95)	7.00 (5.46)
Missing, *n* (%)	1 (0.3%)	0 (0%)	0 (0%)	1 (0.1%)
Baseline HbA1c (%), mean (SD)	8.29 (0.718)	8.31 (0.753)	8.25 (0.661)	8.28 (0.711)
Missing, *n* (%)	1 (0.3%)	3 (1.0%)	0 (0%)	4 (0.4%)
Baseline FPG (mmol/L), mean (SD)	10.6 (2.31)	10.4 (2.11)	10.4 (1.99)	10.5 (2.14)
Baseline eGFR (MDRD, mL/min/1.73m2), mean (SD)	86.8 (18.8)	85.8 (17.5)	88.1 (19.7)	86.9 (18.7)
Baseline C-Peptide (nmol/L), mean (SD)	0.925 (0.356)	0.936 (0.345)	0.920 (0.375)	0.927 (0.359)
Discontinuation of randomized treatment, *n* (%)				
No	281 (94.0%)	288 (95.4%)	298 (97.7%)	867 (95.7%)
Yes	18 (6.0%)	14 (4.6%)	7 (2.3%)	39 (4.3%)
Use of rescue medication, *n* (%)				
No	254 (84.9%)	265 (87.7%)	286 (93.8%)	805 (88.9%)
Yes	45 (15.1%)	37 (12.3%)	19 (6.2%)	101 (11.1%)

TABLE 3 Number of missing HbA1c values by visit and treatment group.

	Visit									
Treatment group	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Dapagliflozin and Saxagliptin	5	2	5	7	5	10	7	7	14	4
Glimeperide	9	8	8	12	10	14	11	15	16	14
Setting post-ICE values to missing										
Dapagliflozin and Saxagliptin	9	9	9	17	14	14	19	27	33	29
Glimeperide	15	16	15	22	16	19	29	49	58	55

the red points both above and below the line reassures us that positivity is not violated, at least for the FPG variable, where the concern is greatest.

The results for the different estimators are presented in Table 4. The estimates for the potential no-ICE outcome under dapagliflozin + saxagliptin are consistent across all the different variations of the methods, except for G-formula 10 and 11 that are somewhat larger in magnitude. The estimates of the mean potential no-ICE outcome under glimeperide have a greater variability across methods, compared to the variability of the estimates for dapagliflozin + saxagliptin, with MMRM, G-formula via MI and G-estimation yielding estimates of lower magnitude compared to MI and IPTW. Here, G-formula 10 and 11 also gave larger estimates. The treatment effect estimates range from −0.071 to −0.189 across the different methods. Compared to the treatment effect estimate from the original published analysis, these are consistently slightly smaller in magnitude, which could be explained by the fact that the dataset contained 33 fewer patients, as noted earlier.

For most methods, including additional covariates did not have much impact on the estimates, either of the mean potential outcomes or the treatment effect. This indicates that the impact of additional confounding due to these covariates was small. For the G-formula, however, there does seem to be an impact of including these additional covariates, particularly for the estimates of the two mean potential outcomes. As it was only with this method, it is unlikely due to confounding, and more likely either due to the modeling choices made for the pooled models used in G-formula, or the different assumptions made by G-formula regarding missing data, as described above.

Comparing methods that only use values before the ICE, IPTW had slightly larger SE as expected. Including values

TABLE 4 Potential outcome mean and treatment effect estimates under no rescue or discontinuation of randomized treatment using the different estimators.

No	Method	Dapagliflozin and Saxagliptin, estimate (SE)	Glimeperide, estimate (SE)	Treatment effect, estimate (SE)	Computational time (mins)
		Using only values before the ICE			
	MMRM				
	Original published analysis	−1.20 (0.05)	−0.99 (0.05)	−0.21 (0.07)	
1	Replicating original analysis	−1.230 (0.0430)	−1.093 (0.0449)	−0.137 (0.0621)	1
	Multiple imputation				
2	HbA1c and treatment	−1.234 (0.0425)	−1.115 (0.0443)	−0.119 (0.0615)	1
3	+ all baseline covariates	−1.234 (0.0421)	−1.115 (0.0445)	−0.119 (0.0611)	1
4	+ all time-varying covariates	−1.234 (0.0420)	−1.120 (0.0439)	−0.114 (0.0604)	4
	IPTW∗				
5	HbA1c and treatment	−1.249 (0.0430)	−1.144 (0.0462)	−0.105 (0.0630)	19
6	+ all baseline covariates	−1.247 (0.0437)	−1.143 (0.0468)	−0.104 (0.0638)	32
7	+ all time-varying covariates	−1.215 (0.0521)	−1.133 (0.0488)	−0.082 (0.0717)	35
8	Separate ICE mechanisms	−1.207 (0.0611)	−1.136 (0.0493)	−0.071 (0.0783)	368
		Exploiting post-ICE values			
	G-formula				
9	HbA1c and treatment	−1.272 (0.0412)	−1.125 (0.0453)	−0.147 (0.0600)	241
10	+ all covariates	−1.433 (0.0433)	−1.246 (0.0465)	−0.187 (0.0627)	1506
11	Separate ICE mechanisms	−1.420 (0.0446)	−1.231 (0.0477)	−0.189 (0.0629)	1561
	G-formula via MI				
12	HbA1c and treatment	−1.205 (0.0457)	−1.062 (0.0533)	−0.143 (0.0637)	2
13	+ all covariates	−1.238 (0.0501)	−1.083 (0.0449)	−0.155 (0.0690)	6
	G-estimation∗				
15	HbA1c and treatment	−1.212 (0.0440)	−1.065 (0.0531)	−0.147 (0.0639)	15
16	+ all covariates	−1.206 (0.0445)	−1.053 (0.0532)	−0.153 (0.0658)	25

∗These methods required a combination of MI and bootstrapping to deal with missing data and derive corresponding standard errors.
As described further in the text, estimation failed for methods 14 and 17.

post-ICE has the potential to improve the efficiency of estimates, at the cost of having to make and rely on more modeling assumptions (Olarte Parra et al., 2022), but in these analyses the SEs of G-formula, G-formula via MI and G-estimation were comparable to those of the estimators which only used data up until the ICE occurred. The treatment effect estimates were larger from the methods that used post-ICE data, but we caution that there is no reason such a systematic difference would be expected in general. Such differences could be due to the modeling assumptions in the methods that used post-ICE data not being correct, but such differences could also occur randomly as a result of the post-ICE data estimators being more efficient. Allowing for distinct mechanisms for the two different ICEs in IPTW increased the SE, as one would expect. For G-formula, allowing for separate ICE mechanisms had relatively little impact on inferences. For G-formula via MI (method 14) and G-estimation (method 17), MI of the missing data accounting for separate ICE mechanisms failed due to sparsity issues, and so estimates for these could not be obtained. Similarly, for G-estimation (methods 15 and 16), in some bootstrap samples estimates of the coefficient of the ICE whose effect was to be removed could not be estimated, and in such cases we set ψ_k=0.

In terms of computational time, MMRM and MI were much faster than the other methods. This is because they handle intermittent missing data in the same process as estimating potential outcomes and they do not require bootstrapping to obtain estimates of the SE. IPTW and G-estimation were much slower because they required handling missing data with MI as a first step and also bootstrapping. Somewhat unexpectedly, the slowest method was G-formula even though we did not combine it with MI to handle missing data. As explained before, the package fits a pooled model across visits so there are less models and parameters to be estimated compared to those methods fitting separate models per visit, such as MI or G-estimation. The computational time was greatly improved when using the alternative G-formula via MI that avoids the need for bootstrapping.

6 DISCUSSION

We have described different treatment effect estimators and their statistical assumptions, to account for rescue medication use and treatment discontinuation through the hypothetical strategy in a diabetes trial. Overall, estimated effects and standard errors were quite similar across all the estimators considered. Although the proportions of patients with an ICE in the trial were non-trivial, from a missing data perspective the proportions of unobserved no-ICE outcomes were small in both arms. As such, one can reasonably anticipate that the different estimators will yield similar inferences, and that it will only be in trials with substantially higher occurrence of ICE where materially different inferences might be obtained.

A key decision in choosing an estimator is whether to use one that exploits post-ICE data or not. Estimators that use such data are generally more or, at least, as efficient as those that do not, but only by making additional statistical assumptions regarding the effects of ICEs on subsequent outcomes (Olarte Parra et al., 2022). Since we do not anticipate material differences in inferences between the estimators unless the proportions with an ICE are quite large, we believe that generally it will be preferable to estimate hypothetical estimands using estimators which do not use post-ICE data. This recommendation is further reinforced by the various issues we encountered when using estimators which use post-ICE data relating to missing data and sparsity. Presently trial analyses targeting a hypothetical estimand to a large extent either use repeated measures models such as MMRM or MI, fitted using pre-ICE data only, whereas IPTW is in our experience rarely used in this context. When there are time-varying confounders that have effects on ICE occurrence and the outcome these should be adjusted for, and MI and IPTW both offer a route to doing this. Specifying the model for the ICE occurrence in IPTW is arguably an easier task than modelling the distributions of all time-varying confounders, as required by MI. However, IPTW cannot readily accommodate missingness in time-varying confounders, which occurs often in practice. As such, we view MI as the most desirable approach for the trial's primary analysis, given its ability to adjust for time-varying confounders and handle missing data prior to the ICE.

An important component of the ICH E9 Addendum is to emphasize that trial analyses should include assessments of sensitivity of results to estimator assumptions (ICH, 2019). If MI is adopted, a natural approach to assess robustness to its modelling assumptions is to compare results with those from IPTW, which relies instead on a model for the ICE occurrence. The other key assumption to assess is the MAR/sequential exchangeability assumption. MI provides a flexible approach to assess sensitivity–imputations of unobserved no-ICE outcomes made under MAR can be successively adjusted, or parameters in the imputation models varied, in a tipping point type analysis (O'Kelly and Ratitch, 2014).

The different estimators considered can be applied using standard software. The package for IPTW is very flexible but cannot accommodate missing values. This limitation of handling missing (intermittent) values can be overcome with MI combined with bootstrapping which is computationally intensive. As already explained the `gfoRmula` package allows to use datasets with missing values (under complete case type assumptions) but still requires bootstrapping. An attractive alternative is G-formula via MI that avoids bootstrapping and is much faster. We recently developed a package to facilitate its implementation (Bartlett et al., 2023).

For methods whose implementation in R use the dataset in wide format (MI, G-formula via MI and G-estimation), separate models are fitted at each visit. These models are more flexible than those using the long form (IPTW and G-formula) because they allow for different covariate effects and intercept per visit. The long form dataset implementations use a more parsimonious model which has the advantage of improved precision but at the expense of potential bias due to model misspecification. Thus, more complex model specifications are required to achieve the same flexibility as the wide format implementations. For example, in the G-formula model for HbA1c, we included randomized treatment and earlier values of HbA1c (the value at the previous visit and the average until the visit) as covariates. An alternative would be to additionally include time or visit as a categorical variable and its interaction with treatment to more flexibly model the evolution of the treatment effects.

In some trials, it will often be the case that the ICE is a (near-)deterministic function of the covariate history, because the ICE corresponds to clinical decisions based on observations on the patient. In such cases, inclusion of such covariates is crucial to ensure the sequential exchangeability assumption is satisfied. This, however, leads to a (near-)violation of the positivity assumption. In this context, IPTW estimators may have large variance, which is a logical consequence of the lack of overlap in the corresponding covariate's distribution between patients with or without an ICE. In contrast, estimators that model covariates and outcomes, such as G-formula, MMRM, and MI, may be more stable because they extrapolate beyond the data based on the model assumptions. Thus, in situations where positivity is violated, these estimators may be preferable to IPTW, provided the modeling assumptions and the extrapolation based on these is deemed reasonable given subject matter knowledge.

When a hypothetical estimand is chosen, we have described how estimation relies on variables which affect outcomes and the ICE being used in the analysis. As such, at the design stage trials targeting hypothetical estimands should identify such variables and ensure as best as possible that they will be measured when the trial is run. Moreover, trial designs and protocols could be modified in order to avoid or minimize violation of the positivity assumption. For example, in the case of rescue treatment in diabetes trials, rescue could be initiated probabilistically, rather than as a deterministic function of biomarkers.

In this paper, we considered the hypothetical estimand that was targeted by the analysis performed from the original trial publication. This corresponds to the effect had all patients remained on randomized treatment without rescue throughout follow-up. To drug sponsors and regulatory agencies it may be of interest to isolate the sole effects of the study drugs, separate from any effects of rescue treatment and discontinuation. For other stakeholders, it may be less interesting, since in practice rescue would not be withheld if it were clinically indicated, and because we would similarly not prevent discontinuation in certain situations, such as if the patient experienced an adverse event linked to the study drugs. They may moreover be somewhat ill-defined unless one can describe how the corresponding hypothetical trial, where all discontinuations are prevented and rescue is withheld, would be run.

If instead both rescue and discontinuation are handled using the treatment policy strategy, outcomes may reflect what would be realized in the presence of these possible events. Such an interpretation requires, however, that rescue use and discontinuation is reflective of what would be seen in routine practice. Moreover, interpretation of such effects is arguably difficult if the ICE occurrence differed between arms. In cases such as the diabetes trial considered here, a less efficacious drug may lead to more rescue use, which leads to improved short term outcomes, but does not provide a long term viable treatment solution for the patient. A treatment policy analysis would then suggest such a drug is preferable, but only because its use led to more rescue treatment being administered (Keene et al., 2021).

An alternative approach is to recognize the ICEs in the endpoint definition, by using the composite strategy. While this may be straightforward for binary outcomes, it is more difficult for continuous outcomes, like here, since there is no natural or obvious value to assign if a patient has an ICE. An alternative approach that may be attractive is to use a win ratio or "proportion in favour of treatment" type approach (Buyse, 2010). This requires explicit consideration of what combinations of outcomes and ICEs constitute better responses, which must be made with clinical input. While the resulting effect measure is no longer on the original outcome variable scale, such estimands avoid conception of hypotheticals and can properly recognize that certain ICEs (eg, rescue treatment use) constitute "bad" outcomes.

All of these considerations regarding choosing the strategy to deal with a particular ICE resemble the considerations of the target trial emulation framework (Hernán and Robins, 2016). With this framework, observational studies are analyzed to estimate the treatment effect in an ideal trial. For trials, it may be unethical to randomly assign insulin as rescue for patients with high glucose levels, but it is possible to imagine a hypothetical trial where this could be the case. With such a target trial in mind, the target estimand can be described.

We hope that these considerations of the suitability of a hypothetical strategy to handle a particular ICE in a given context, the step-by-step description of different available estimators and further considerations of the implications of their different underlying assumptions are useful for planning, conducting and analyzing trials using the estimand framework.

ACKNOWLEDGMENTS

D.W. is a full time employee of AstraZeneca and owns shares in AstraZeneca and provided some of his time to support COP in her research. This publication is based on research using data from data contributors AstraZeneca that has been made available through Vivli, Inc. Vivli has not contributed to or approved, and is not in any way responsible for, the contents of this publication. J.W.B.'s past and present institutions have received consultancy fees for his advice on statistical methodology from AstraZeneca, Bayer, Novartis, and Roche. J.W.B. has received consultancy fees from Bayer and Roche.

SUPPLEMENTARY MATERIALS

Supplementary material is available at *Biometrics* online.

Web Appendices, Figures and code referenced in Sections 4 and 5 are available with this paper at the Biometrics website on Oxford Academic.

FUNDING

This work was funded by UK Medical Research Council grant MR/T023953/1 and MR/T023953/2.

CONFLICT OF INTEREST

DW is a full time employee of AstraZeneca and owns shares in AstraZeneca and provided some of his time to support COP in her research. This publication is based on research using data from data contributors AstraZeneca that has been made available through Vivli, Inc. Vivli has not contributed to or approved, and is not in any way responsible for, the contents of this publication. JB's past and present institutions have received consultancy fees for his advice on statistical methodology from AstraZeneca, Bayer, Novartis, and Roche. JB has received consultancy fees from Bayer and Roche.

DATA AVAILABILITY

The data that support the findings in this paper may be obtained in accordance with AstraZeneca's data sharing policy described at https://astrazenecagrouptrials.pharmacm.com/ST/Submission/Disclosure. AstraZeneca Vivli member page is also available outlining further details: https://vivli.org/ourmember/astrazeneca/.

REFERENCES

Bartlett, J. W., Olarte Parra, C., Granger, E., Keogh, R. H., van Zwet, E. W. and Daniel, R. M. (2023). G-formula for causal inference via multiple imputation. arXiv, arXiv:2301.12026, preprint.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J. and Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163, 1149–1156.

Buyse, M. (2010). Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in Medicine*, 29, 3245–3257.

Hernán, M. A. and Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183, 758–764.

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. United States: Boca Raton: Chapman & Hall/CRC, https://miguelhernan.org/whatifbook (Accessed on 25/04/2021).

Holzhauer, B., Akacha, M. and Bermann, G. (2015). Choice of estimand and analysis methods in diabetes trials with rescue medication. *Pharmaceutical Statistics*, 14, 433–447.

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K. and Sterne, J. A. (2014). Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, 14, 1–10.

ICH, (2019). *International Council for Hamonisation Topic E9(R1) on Estimands and Sensitivity Analysis in Clinical Trials*. Available at www.ich.org (Accessed on 10/02/2021).

Keene, O. N., Wright, D., Phillips, A. and Wright, M. (2021). Why ITT analysis is not always the answer for estimating treatment effects in clinical trials. *Contemporary Clinical Trials*, 108, 106494.

Lasch, F. and Guizzaro, L. (2022). Estimators for handling COVID-19-related intercurrent events with a hypothetical strategy. *Pharmaceutical Statistics*, 21, 1258–1280.

Lasch, F., Guizzaro, L., Pétavy, F. and Gallo, C. (2022). A simulation study on the estimation of the effect in the hypothetical scenario of no use of symptomatic treatment in trials for disease-modifying agents for Alzheimer's disease. *Statistics in Biopharmaceutical Research*, 15(2), 386–399.

Leyrat, C., Seaman, S. R., White, I. R., Douglas, I., Smeeth, L., Kim, J. et al. (2019). Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Statistical Methods in Medical Research*, 28, 3–19.

Lipkovich, I., Ratitch, B. and Mallinckrodt, C. H. (2020). Causal inference and estimands in clinical trials. *Statistics in Biopharmaceutical Research*, 12, 54–67.

Loh, W. W., Moerkerke, B., Loeys, T., Poppe, L., Crombez, G. and Vansteelandt, S. (2020). Estimation of controlled direct effects in longitudinal mediation analyses with latent variables in randomized studies. *Multivariate Behavioral Research*, 55, 763–785.

Müller-Wieland, D., Kellerer, M., Cypryk, K., Skripova, D., Rohwedder, K., Johnsson, E. et al. (2018). Efficacy and safety of dapagliflozin or dapagliflozin plus saxagliptin versus glimepiride as add-on to metformin in patients with type 2 diabetes. *Diabetes, Obesity and Metabolism*, 20, 2598–2607.

O'Kelly, M. and Ratitch, B. (2014). *Clinical Trials with Missing Data: A Guide for Practitioners*. United Kingdom: John Wiley and Sons.

Olarte Parra, C., Daniel, R. M. and Bartlett, J. W. (2022). Hypothetical estimands in clinical trials: a unification of causal inference and missing data methods. *Statistics in Biopharmaceutical Research*, 15(2), 421–432.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75, 591–593.

Schomaker, M. and Heumann, C. (2018). Bootstrap inference when using multiple imputation. *Statistics in Medicine*, 37, 2252–2266.

van der Wal, W. M. and Geskus, R. B. (2011). ipw: an R package for inverse probability weighting. *Journal of Statistical Software*, 43, 1–23.

VanderWeele, T. J. and Hernan, M. A. (2013). Causal inference under multiple versions of treatment. *Journal of Causal Inference*, 1, 1–20.

Vansteelandt, S. and Sjolander, A. (2016). Revisiting g-estimation of the effect of a time-varying exposure subject to time-varying confounding. *Epidemiologic Methods*, 5, 37–56.

Wang, N. and Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85, 935–948.

White, I. R. and Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, 29, 2920–2931.

Received: September 1, 2023; Revised: October 19, 2024; Accepted: January 3, 2025