G-formula for causal inference using synthetic multiple imputation

Jonathan W. Bartlett

Department of Medical Statistics
London School of Hygiene & Tropical Medicine
www.thestatsgeek.com

BMS-ANed seminar
4th November 2025

Acknowledgements

This is joint work with:

- Camila Olarte Parra (formerly Bath, LSHTM, now Karolinska)
- Rhian Daniel (Cardiff)
- Emily Granger and Ruth Keogh (LSHTM)
- Erik van Zwet (Leiden)

This work was supported by a UK Medical Research Council grant (MR/T023953/1 and MR/T023953/2).

Motivation

Time-varying treatments and confounders, and the G-formula

Multiple imputation for missing data

G-formula via multiple imputation with complete data

G-formula via MI with missing data

Discussion

Outline

Motivation

In many settings we are interested in effects of time-varying treatments or exposures.

Our exploration of this area stems from analysis of an RCT in diabetes where patients may in addition to randomised treatment, take rescue treatment, or discontinue randomised treatment.

Thus treatment(s) received by patients varied over time.

We were interested in estimating effects of randomised treatment in the hypothetical scenario where patients remain on randomised treatment throughout follow-up and are not given rescue treatment.

There are of course also many settings where treatment or exposure varies over time and in which there is no randomisation.

Motivation

To estimate effects of treatment strategies/sequences, we must consider confounding.

Not just baseline confounders, but also time-varying confounders.

To do this 'correctly' requires use of so called G-methods, developed by Jamie Robins and co-workers [3, 2].

This talk will be about one of these methods **G-formula** (or sometimes G-computation).

We wanted to apply G-formula, but the dataset had missing values.

How to handle the missing values in the context of a G-formula analysis?

Outline

Time-varying treatments and confounders

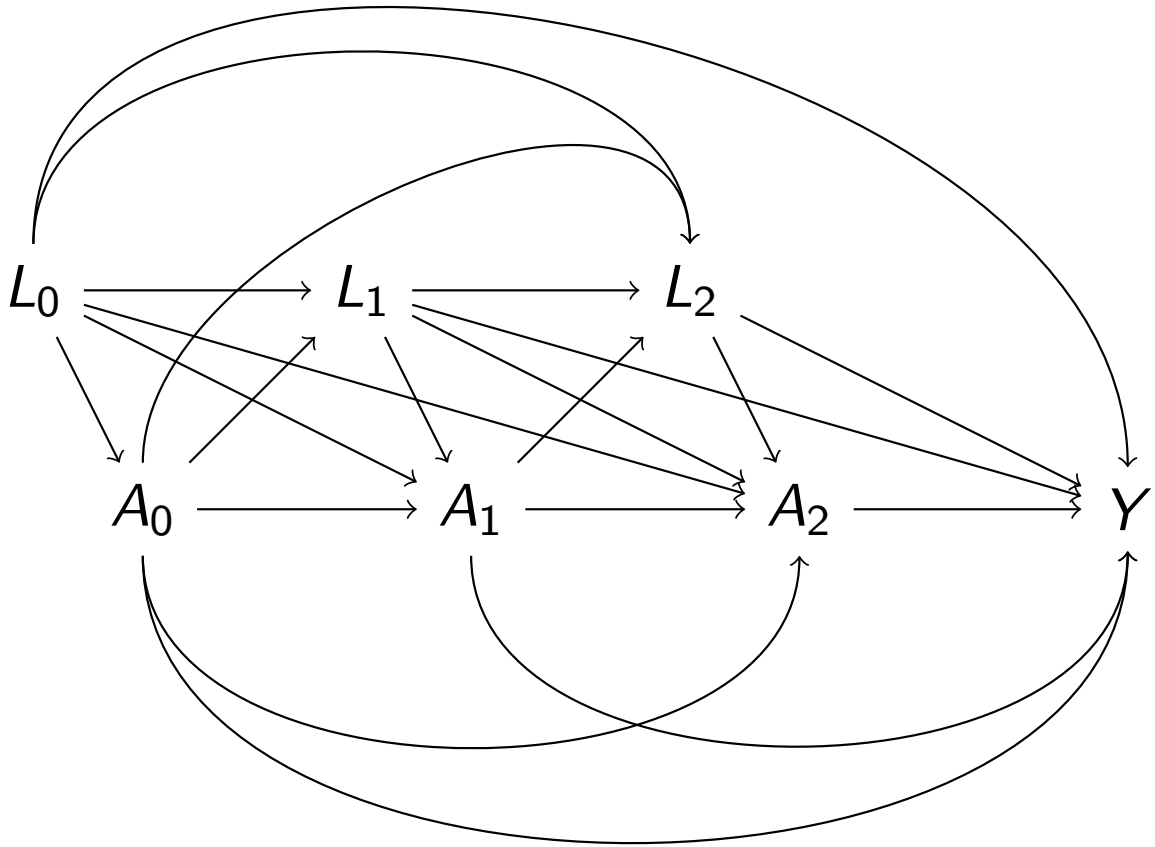
The setting under consideration is the 'standard' time-varying treatment and confounding setup.

A_k denotes treatment at time/visit k, $k = 0, \ldots, K$.

L_k denotes time-varying confounders at visit k.

Y denotes the final outcome of interest.

Directed acyclic graph (DAG)



Potential outcomes and estimands

Let $Y^{a_0,a_1,a_2} = Y^{\bar{a}}$ denote potential outcome if treatment is set to $\bar{a} = (a_0, a_1, a_2)$.

Causal estimands are then contrasts of aspects of the distributions of $Y^{\bar{a}}$ for different values of $\bar{a}$.

For example, the effect of treatment at all times versus no treatment is

$$E(Y^{1,1,1}) - E(Y^{0,0,0})$$

Identification assumptions

Consistency: Interventions on treatment/exposure well defined so that we can assume $Y = Y^{\bar{a}}$ if $\bar{A} = \bar{a}$.

Conditional exchangeability:

$$Y^{\bar{a}} \perp\!\!\!\perp A_k | \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_k \text{ for } k = 0, 1, \ldots, K$$

Positivity: If $f(\bar{a}_{k-1}, \bar{l}_k) > 0$ then

$$P(A_k = a_k | \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_k = \bar{l}_k) > 0$$

G-formula

In G-formula, estimation of $E(Y^{\bar{a}}) = E(Y^{a_0,a_1,a_2})$ is based on

$$E(Y^{\bar{a}}) = \int_{l_0} \int_{l_1} \int_{l_2} E(Y|a_0, a_1, a_2, l_0, l_1, l_2) f(l_2|a_0, a_1, l_0, l_1) f(l_1|a_0, l_0) f(l_0) dl_2 dl_1 dl_0$$

This requires we specify and fit models for

- $f(L_0)$ (in fact, we typically empirically average across this, avoiding need for a model)
- $f(L_1|A_0, L_0)$
- $f(L_2|A_0, A_1, L_0, L_1)$
- $f(Y|A_0, A_1, A_2, L_0, L_1, L_2)$ (in fact, all we need is a model for $E(Y|A_0, A_1, A_2, L_0, L_1, L_2)$)

In the above we have separate models for time-varying confounders at each time point.

In fact, implementations (e.g. `gfoRmula` in R) fit pooled long data form models for time-varying confounders.

G-formula by Monte-Carlo integration/simulation

The integrals are generally analytically intractable.

Thus software implementations use Monte-Carlo integration/simulation.

To estimate $E(Y^{a_0,a_1,a_2})$, based on fitted models, for every individual we:

- simulate L_0^* from $f(L_0)$ (or just use original, i.e. $L_0^* = L_0$)
- simulate L_1^* from $f(L_1|A_0 = a_0, L_0^*)$
- simulate L_2^* from $f(L_2|A_0 = a_0, A_1 = a_1, L_0^*, L_1^*)$
- simulate Y^* from $f(Y|A_0 = a_0, A_1 = a_1, A_2 = a_2, L_0^*, L_1^*, L_2^*)$

  (or just calculate $E(Y|A_0 = a_0, A_1 = a_1, A_2 = a_2, L_0^*, L_1^*, L_2^*)$
- calculate mean of Y^* across individuals

  (or average $E(Y|A_0 = a_0, A_1 = a_1, A_2 = a_2, L_0^*, L_1^*, L_2^*)$ across individuals)

G-formula inference

Inference for such a G-formula estimator is usually performed using bootstrapping:

- take B bootstrap samples of original data
- apply G-formula to each bootstrap sample
- construct bootstrap SEs / confidence intervals

G-formula with missing data

What if we have some missing data?

- `gformula` in Stata imputes missing values once stochastically
- `gfoRmula` in R fits models for time-varying confounders using 'complete cases'

We wanted to explore how multiple imputation (MI) could be used to impute missing data, followed by G-formula for estimation of causal effects.

Outline

Multiple imputation

- Multiple imputation (MI) is a flexible and now very popular approach for handling missing data.
- It relies (at least in its usual form) on assuming data are missing at random.
- Here I will briefly introduce it in a simple setting with one variable containing missing values.

Parametric imputation model

Let Y contain missing values and X be fully observed.

We want to generate imputations of missing values of Y that

- have the same distribution as the original values had they been observed, and
- have the 'right' relationship with X.

To do this we specify a parametric imputation model for $f(Y \mid X, \psi)$ with parameters ψ.

E.g. a linear model:

$$Y = \alpha + \beta X + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

so here $\psi = (\alpha, \beta, \sigma^2)$.

Single stochastic imputation

If Y is MAR given X, fitting $f(Y \mid X, \psi)$ to those with Y observed gives unbiased estimates $\hat{\psi} = (\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$.

We can then draw (randomly) from $f(Y \mid X, \hat{\psi})$ to impute the missing values.

For the linear regression imputation model, for each missing Y draw using

$$Y^* = \hat{\alpha} + \hat{\beta}X + \varepsilon^*, \qquad \varepsilon^* \sim \mathcal{N}(0, \hat{\sigma}^2).$$

The imputed dataset can then be analysed using the analysis of interest.

Single vs. multiple imputation

There is nothing wrong with single stochastic imputation for point estimation.

There are two problems however:

1. It is somewhat inefficient statistically, because of Monte-Carlo error. Repeating the process (multiple imputation) and averaging our final estimates reduces this.
2. Naive standard errors calculated on the imputed dataset(s) ignore uncertainty due to the imputation process.

There are various ways of solving 2. Rubin's *Bayesian* MI is one convenient route.

Bayesian MI and priors

Rubin's MI is by construction **Bayesian**, but builds on the single stochastic imputation method.

- Specify a prior distribution for the imputation model parameters $f(\psi)$.
- In the Bayesian paradigm, both parameters and missing values are treated as random variables.
- The prior $f(\psi)$ encodes prior belief/evidence about ψ before we see the data.
- Implementations of MI are often based on so-called 'non-informative' priors.

Bayesian predictive distribution

Let $\mathbf{Y}^{\text{mis}}$ and $\mathbf{Y}^{\text{obs}}$ denote missing and observed values of Y, and $\mathbf{X}$ the (fully observed) X values.

To create imputations of $\mathbf{Y}^{\text{mis}}$ perform two steps:

1. Draw $\tilde{\psi}$ from the posterior $f(\psi \mid \mathbf{Y}^{\text{obs}}, \mathbf{X})$.
2. Impute $\mathbf{Y}^{\text{mis}}$ from $f(\mathbf{Y}^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \mathbf{X}, \tilde{\psi})$.

Repeat these two steps M times to create M imputed datasets.

Step 1 is critical for Rubin's simple combination rules to be valid.

Bayesian imputation for linear regression

Consider the linear regression imputation model $Y = \alpha + \beta X + \varepsilon$.

1. Draw $\tilde{\psi} = (\tilde{\alpha}, \tilde{\beta}, \tilde{\sigma}^2)$ from $f(\psi \mid \mathbf{Y}^{\text{obs}}, \mathbf{X})$.
   With one variable missing, this step only depends on cases with Y observed.
2. For *each* missing Y, impute

$$Y^* = \tilde{\alpha} + \tilde{\beta} X + \varepsilon^*, \qquad \varepsilon^* \sim \mathcal{N}(0, \tilde{\sigma}^2),$$

   using that individual's X value.

This creates the first imputed dataset. Repeat to create datasets $2, 3, 4, \ldots, M$.

Bayesian imputation using other models

Imputation can also be performed using other regression models suited to the variable being imputed.

- e.g. logistic regression for binary variables.

The principles remain the same; only steps 1 and 2 change depending on the model.

For more on different imputation methods in R see `?mice::mice`.

The method obviously extends to the case where $\mathbf{X} = (X_1, X_2, \ldots, X_p)$.

It can also be extended to impute missing values in multiple variables (e.g. via the chained equations method).

Analysing imputed datasets (Rubin's rules)

We have created M imputed datasets that are now complete. How do we analyse them?

Assume interest is in a parameter θ, e.g.:

- the mean or median of a variable
- a proportion in a particular category (level) of a factor variable
- a coefficient in a regression model

The approach is often referred to as 'Rubin's rules'.

Step 1: analyse each imputation separately

Perform the analysis (e.g. estimate the mean or fit the regression) on *each* imputed dataset separately.

From each we need the estimate $\hat{\theta}_m$ and its variance $\widehat{\mathrm{Var}}(\hat{\theta}_m)$.

Imputation	Estimate	Variance
1	3.9	0.25
2	4.3	0.31
...	...	...
M	4.5	0.29

Step 2: combine the estimates

Pool the estimates using Rubin's rules.

The MI estimate of θ is

$$\hat{\theta}_{\text{MI}} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m,$$

where $\hat{\theta}_m$ is the estimate from the mth imputed dataset.

Step 3: between-imputation variance

To estimate the variance of $\hat{\theta}_{MI}$ we **don't** simply average the variances from each dataset.

First compute the between-imputation variance:

$$\hat{\sigma}_b^2 = \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\theta}_m - \hat{\theta}_{\text{MI}} \right)^2 .$$

This is the usual unbiased sample variance of $\hat{\theta}_1, \ldots, \hat{\theta}_M$.

$\hat{\sigma}_b^2$ measures how much estimates of θ vary across imputations, i.e. uncertainty due to missing data.

Step 4: within-imputation variance

Next compute the within-imputation variance:

$$\hat{\sigma}_w^2 = \frac{1}{M} \sum_{m=1}^{M} \widehat{\mathrm{Var}}(\hat{\theta}_m),$$

where $\widehat{\mathrm{Var}}(\hat{\theta}_m)$ is the estimated variance from the mth analysis.

$\hat{\sigma}_w^2$ measures uncertainty from finite sample size (the usual source of sampling variability).

Step 5: variance of $\hat{\theta}_{\text{MI}}$

Estimate the variance of $\hat{\theta}_{\text{MI}}$ by

$$\hat{\sigma}^2_{\text{MI}} = \left(1 + \frac{1}{M}\right) \hat{\sigma}^2_b + \hat{\sigma}^2_w.$$

So the uncertainty in $\hat{\theta}_{\text{MI}}$ is (essentially) the sum of between and within imputation variances.

MI and G-formula

Could we use MI to implement the G-formula method?

We could:

- use MI to generate (e.g. $M = 10$) completed datasets
- apply G-formula to each of the M datasets, yielding a point estimate and SE
- pool estimates and SEs using Rubin's rules

If bootstrapping is used for SE estimation in each imputed dataset, the overall procedure is quite computationally costly!

Is there a faster alternative, avoiding bootstrapping?

G-formula and imputation

For those (like me) more familar with missing data methods, G-formula can be seen as a form of single stochastic imputation of the longitudinal history under the treatment regime of interest.

In fact, to reduce Monte-Carlo error, implementations of G-formula create multiple imputations of these, and then average the imputed Y^* across individuals and across imputations.

The close links between MI and G-formula by simulation begs the question - could we use Rubin's combination rules, rather than bootstrapping, for inference?

Outline

G-formula via multiple imputation - earlier work

Westreich *et al* previously highlighted close connections between G-formula and MI in single time point setting [8]:

L	A	Y^0	Y^1
2.5	0	3.4	NA
7.3	1	NA	4.5
4.6	1	NA	5.7
4.2	0	4.2	NA

Impute Y^0 for those with $A = 1$, using L. Impute Y^1 for those with $A = 0$, using L.

Calculate difference in means of Y^1 and Y^0 in imputed datasets.

Westreich *et al* stated that Rubin's variance estimator cannot be used because each individual contributes to both treated and untreated calculations.

G-formula via multiple imputation - earlier work

Here individual's observed outcome is retained.

In the longitudinal setting, G-formula simulates confounders and outcomes for all individuals afresh.

Indeed, in this setting, it may be that no individuals followed precisely the treatment regime of interest.

We can nonetheless use MI to implement G-formula, exploiting existing work on using MI to generate synthetic datasets.

G-formula via MI

For the longitudinal setup earlier, we can use MI to estimate $\mu = E(Y^{\bar{a}})$ in a G-formula type approach by:

1. Augment observed data with additional n_{syn} rows, setting L_0, L_1, L_2, Y to missing in the augmented rows to missing, and A_0, A_1, A_2 to value $\bar{a} = (a_0, a_1, a_2)$.
2. Run MI on the augmented dataset, generating M imputations.
3. For imputation m ($m = 1, \ldots, M$), calculate mean of Y from the augmented part of the dataset.
4. Average estimated means across M imputations (denoted $\hat{\mu}$) as estimator of $\mu = E(Y^{\bar{a}})$.

G-formula via MI - data structure

E.g. data structure for $\bar{a} = (1, 1, 1)$ is

R	L_0	A_0	L_1	A_1	L_2	A_2	Y
1	-0.3	0	0.5	0	2.2	1	1.3
1	2.3	1	4.2	1	4.6	1	5.5
1	-0.5	1	0.4	0	0.8	1	1.9
0	NA	1	NA	1	NA	1	NA
0	NA	1	NA	1	NA	1	NA
0	NA	1	NA	1	NA	1	NA

$R = 1$ indicates originally observed data

$R = 0$ indicates augmented data

G-formula via MI - implementation details

We have a block monotone missingness pattern in the augmented dataset.

Due to our earlier model assumptions, we can impute sequentially moving forwards in time:

1. Impute L^0
2. Impute $L_1|A_0, L_0$
3. Impute $L_2|A_0, A_1, L_0, L_1$
4. Impute $Y|A_0, A_1, A_2, L_0, L_1, L_2$

This means if we use for example chained equations MI software, there is no need to iterate around models.

We specify imputation equations as per above, and set iterations to 1.

Contrasts of treatment regimes

In practice we are interested in contrasts of the form $E(^{\bar{a}_1}) - E(^{\bar{a}_2})$ for regimes $\bar{a}_1$ and $\bar{a}_2$.

To estimate this, add augmented rows with $\bar{A} = \bar{a}_1$ and another set with $\bar{A} = \bar{a}_2$.

In the imputed datasets, calculate difference in sample means.

Inference for G-formula via MI estimator

How to estimate $\text{Var}(\hat{\mu})$ and conduct inference?

Ordinarily with MI we use Rubin's rules.

Estimate variance in each imputation and average these, yielding within-imputation variance $\hat{V}$.

Estimate variance of estimated means across M imputations, yielding between-imputation variance $\hat{B}$.

Then $\widehat{\text{Var}}(\hat{\mu}) = (1 + M^{-1})\hat{B} + \hat{V}$.

Unfortunately this does not work here - Rubin's variance estimator is much larger than the true $\text{Var}(\hat{\mu})$.

This is due to a form of uncongeniality - the imputation and analysis models are being fitted to different portions of the dataset.

Multiple imputation for synthetic samples/populations

Within the survey sampling field, there is an established literature on using MI to impute partially or fully synthetic datasets.

The motivation here is concern over confidentiality if survey data were released to analysts.

MI is used to impute/simulate variables for new/synthetic individuals, ensuring confidentiality of original participants.

Variance estimation for synthetic MI

An unbiased variance estimator developed by Raghunathan, Reiter and Rubin [5] for this synthetic MI setting is

$$(1 + M^{-1})\hat{B} - \hat{V}.$$

In our paper, we use asymptotic theory for MI estimators of Robins and Wang [6] to show that the above variance estimator is asymptotically unbiased for the 'G-formula via MI' estimator.

Simulation setup

To evaluate G-formula via MI approach, we performed simulations.

$n_{\text{obs}} = n_{\text{syn}} = 500$

$10,000$ simulations per scenario.

We simulated with two intermediate follow-ups and a final outcome Y.

Sequential imputation using `mice` package in R, with M synthetic imputations.

Simulation setup

$$
\begin{aligned}
L_0 &\sim N(0,1) \\
P(A_0 = 1|L_0) &= \text{expit}(L_0) \\
L_1 &\sim N(A_0 + L_0, 1) \\
P(A_1 = 1|A_0, L_0, L_1) &= \text{expit}(A_0 + L_1) \\
L_2 &\sim N(A_1 + L_1, 1) \\
P(A_2 = 1|A_0, A_1, L_0, L_1, L_2) &= \text{expit}(A_1 + L_2) \\
Y &\sim N(A_2 + L_2, 1)
\end{aligned}
$$

We target $E(Y^{1,1,1}) - E(Y^{0,0,0})$, which has true value 3.

Number of imputations M

The variance estimator $(1 + M^{-1})\hat{B} - \hat{V}$ can be negative, due to noise in $\hat{B}$ as estimate of true between-imputation variance.

To examine how large M needs to be to avoid this, we evaluated $M = \{5, 10, 25, 50, 100\}$.

If on a given dataset the estimated variance was negative, we added new sets of M imputations until it became non-negative.

We report the mean and max. value of M required across the $10,000$ simulations.

Simulation results

M	Bias	Emp. SE	Est. SE	95% CI	Mean M	Max M
5	-0.002	0.242	0.236	99.4	6.2	25
10	0.001	0.229	0.223	98.4	10.4	30
25	0.000	0.223	0.220	95.6	25.0	50
50	0.000	0.217	0.219	95.2	50.0	50
100	0.004	0.218	0.219	95.0	100.0	100

- Estimates are unbiased for true effect ($= 3$).
- For $M \geq 25$ 95% coverage is reasonable.
- Negative variance issue is rare with M as low as 25, and never occurred with $M \geq 50$.

Interim conclusions

MI provides a potentially convenient route to performing G-formula without need for bootstrapping.

Inference seems reliable for M as low as 25, which is computationally far fewer than number of bootstraps typically used (e.g. $1,000$).

Fairly easy to implement manually using `mice` in R, but we also provide a package `gFormulaMI` to facilitate the method.

Outline

G-formula and missing data

Our original motivation was how to combine MI to handle missing data with G-formula to estimate the causal effects of interest.

Given we've seen that MI can be used to perform G-formula when data are complete, can we use it to impute any missing data as well?

Yes. We impute the combination of the missing actual data and the missing potential outcome data (in the augmented part).

Imputing missing data with MI G-formula

There are (at least) two approaches:

1. Impute missing actual data and missing potential outcome data in one go.
2. First impute missing actual data, then impute missing potential outcome data conditional on these imputations.

We believe option 2 is more attractive.

We only have to impute the (usually) small amount of missing actual data to create a monotone pattern. The remaining missing (potential outcome) values are then imputed as before, with no iterations required at the second stage.

This approach is well established in the context of using MI with longitudinal data where the missingness pattern is almost monotone.

Simulation setup

To the previous setup, we made some values in L_1, A_1, L_2, A_2, Y independently missing completely at random.

We varied the probability $p = \{0.05, 0.1, 0.25, 0.5\}$ that values in each were missing. Note p refers to the marginal probability that values in **each** variable are missing.

We imputed missing values $M = 50$ using `mice` with default settings.

For each imputed dataset, we then augmented as described previously, and imputed missing potential outcomes once as described earlier.

Inferences were then again based on Raghunathan's variance estimator.

Simulation results

π	Bias	Emp. SE	Mean est. SE	95% CI
0.05	-0.001	0.225	0.224	95.4
0.10	-0.003	0.231	0.231	95.3
0.25	-0.008	0.259	0.258	95.4
0.50	-0.011	0.360	0.361	95.0

- As expected, estimator becomes more variable with increasing missingness.
- Raghunathan variance estimator and 95% CI performing well.

Outline

Discussion

- MI seems like an attractive route to implementing G-formula, particularly when some data are missing.
- Exploits connections between causal inference and missing data and software developed for the latter.
- See our paper [1] for more details.
- Our R package `gFormulaMI` which utilises `mice`, is available on CRAN.
- Application of the approach for the diabetes RCT mentioned at the start is reported in [4].

Discussion

I have not mentioned various important things, including:

- The 'complete case' assumption made by `gfoRmula` may be plausible than the MAR assumption relied upon by MI.
- Often a large portion of missingness is due to dropout/censoring, and this can be viewed as part of the time-varying treatment [2]. Nevertheless, sporadic/intermittent missingness is often also present.
- Iterative conditional expectation (ICE) version of G-formula avoids Monte-Carlo simulation/integration [7].
- ICE moreover permits variance estimation using estimating equation theory, avoiding the need for bootstrap [9].

References I

[1] J. W. Bartlett, C. Olarte Parra, E. Granger, R. H. Keogh, E. W. van Zwet, and R. M. Daniel.

G-formula with multiple imputation for causal inference with incomplete data.

Statistical Methods in Medical Research, 34:1130–1143, 2025.

[2] M. Hernán and J. Robins.

Causal Inference: What If.

Boca Raton: Chapman & Hall/CRC, 2020.

[3] A. I. Naimi, S. R. Cole, and E. H. Kennedy.

An introduction to g methods.

International Journal of Epidemiology, 46(2):756–762, 2017.

References II

[4] C. Olarte Parra, R. M. Daniel, D. Wright, and J. W. Bartlett.

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

Biometrics, 81(1):ujae167, 2025.

[5] T. E. Raghunathan, J. P. Reiter, and D. B. Rubin.

Multiple imputation for statistical disclosure limitation.

Journal of Official Statistics, 19(1):1, 2003.

[6] J. M. Robins and N. Wang.

Inference for imputation estimators.

Biometrika, 85:113–124, 2000.

[7] L. Wen, J. G. Young, J. M. Robins, and M. A. Hernán.

Parametric g-formula implementations for causal survival analyses.

Biometrics, 77(2):740–753, 2021.

References III

[8] D. Westreich, J. K. Edwards, S. R. Cole, R. W. Platt, S. L. Mumford, and E. F. Schisterman.

Imputation approaches for potential outcomes in causal inference.

International Journal of Epidemiology, 44(5):1731–1737, 2015.

[9] P. N. Zivich, R. K. Ross, B. E. Shook-Sa, S. R. Cole, and J. K. Edwards.

Empirical sandwich variance estimator for iterated conditional expectation g-computation.

Statistics in Medicine, 43(29):5562–5572, 2024.