Hypothetical estimands – a unification of causal inference and missing data methods

Jonathan W. Bartlett

Department of Medical Statistics
LSHTM
https://thestatsgeek.com

Addressing intercurrent events
Treatment policy and hypothetical strategies
Joint EFSPI & BBS virtual event
15th December 2022

Acknowledgements

This is joint work with Camila Olarte Parra (LSHTM) and Rhian Daniel (Cardiff).

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

Thanks to David Wright for his and AstraZeneca's input, as an industrial collaborator on the project.

Causal inference and G-formula

Missing data approaches

Conclusions

Outline

Causal inference and G-formula

Missing data approaches

Conclusions

Causal inference with time-varying treatment

Causal inference is well developed for estimating effects of time-varying treatments.

Here a key issue is time-varying confounding.

Correctly handling the latter requires the use of special (G-) methods, mostly developed by James Robins & coworkers.

We can embed the occurrence of ICEs into this framework by treating occurrence of the ICEs as a time-varying treatment.

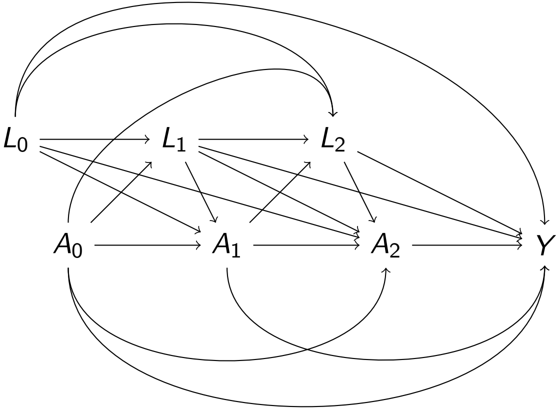
Notation

Think of a diabetes RCT with HbA1c as outcome.

- Randomised treatment A_0
- Occurrence of ICE at time $t > 0$, A_t (e.g. receiving rescue treatment or discontinuation of randomised treatment)
- Outcome of interest Y (e.g. HbA1c at final time point)
- Common causes of ICEs and outcome L_t (e.g. HbA1c and fasting plasma glucose (FPG) measured at time t)

For simplicity, in the following I will assume we have just two follow-up time points at which ICE could occur.

Directed acyclic graph (DAG)



Potential outcomes and hypothetical estimand

Let Y^{a_0,a_1,a_2} denote potential outcome if treatment A_0 is set to value a_0, ICE A_1 is **set** to a_1, and ICE A_2 is **set** to a_2.

The (**whoops, 'a'**) hypothetical estimand then targets

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

In words: the mean difference in outcome between treatments if we prevented ICE from occurring at any time.

Assumptions - consistency

Consistency (not usual ‘statistical consistency’)

Interventions to prevent ICE are well defined so that $Y = Y^{a_0,a_1,a_2}$ if $A_0 = a_0$, $A_1 = a_1$, $A_2 = a_2$

$\Rightarrow$ in the **actual trial**, for a patient who did not need rescue or discontinue treatment, their actual outcome Y is identical to the outcome they would have in the hypothetical trial where we withhold rescue and prevent discontinuation.

To convincingly argue why consistency would hold, we need to to try and articulate how the ICEs would be prevented.

C.f. Hernán [2] on causal effect of obesity on mortality being ill-defined - effect depends on how you change someone’s weight.

Assumptions - no unmeasured confounding

Conditional exchangeability (no unmeasured confounding)

In our case, this means that ICE occurrence at a given visit is independent of final outcome, conditional on measured past.

This holds under the DAG shown previously.

But, we need to measure (and adjust for in the analysis) **all common causes of ICE and outcome** Y.

In a diabetes trial, this means we should adjust for FPG, not just HbA1c, if FPG influences rescue decisions (c.f. Holzhauer *et al* [3])

Assumptions - positivity

Positivity

At all values of L_0 and L_1 which can occur, there is a non-zero probability of the ICE A_1 **not occurring** (similar for A_2).

This would be violated if rescue treatment A_1 is initiated deterministically based on L_1.

This could happen with insulin rescue in diabetes trials, if patients get rescued if and only if FPG exceeds a threshold.

Positivity is not actually needed for G-formula, but then we are relying on the model to extrapolate beyond the data.

Estimation

To estimate $E(Y^{0,0,0})$ and $E(Y^{1,0,0})$, we can use:

- G-formula (sometimes called G-computation)
- Inverse probability of treatment weighting (here 'treatment' is A_0, A_1, A_2)
- G-estimation (see Florian's talk)

I will focus on G-formula, and how it relates to MMRM and multiple imputation.

G-formula v1 (‘standard version’)

To estimate $E(Y^{0,0,0})$:

1. specify and fit models for
   - $f(L_1|A_0, L_0)$
   - $f(L_2|A_0, A_1, L_0, L_1)$
   - $f(Y|A_0, A_1, A_2, L_0, L_1, L_2)$
2. for every patient
   - simulate L_1^* from $f(L_1|A_0 = 0, L_0)$
   - simulate L_2^* from $f(L_2|A_0 = 0, A_1 = 0, L_0, L_1^*)$
   - simulate Y^* from $f(Y|A_0 = 0, A_1 = 0, A_2 = 0, L_0, L_1^*, L_2^*)$
   - calculate mean of Y^* across patients

For $E(Y^{1,0,0})$ replace $A_0 = 0$ with $A_0 = 1$ in the second part.

G-formula intuition and points to note

G-formula simulates (imputes!) longitudinal history (L_1, L_2, Y) for every patient under the hypothetical scenario of interest where ICE does not occur.

Observations of L_2 and Y after occurrence of ICE in the real trial are (by default) not excluded from the model fitting process.

But this requires us to model the effects of ICE occurrence (effects of rescue/discontinuation) on L_2 and Y.

This differs to a 'standard' MMRM analysis, which discards post-ICE data.

If trial did not collect data after ICE, we of course cannot model what happens post-ICE.

G-formula v2 - excluding data after ICE

In fact, since for the hypothetical estimand we are only interested in no ICE potential outcomes, we can avoid modelling effects of ICE A_1 on L_2 and Y and A_2 on Y.

We can specify models for:

- $f(L_1|A_0, L_0)$ (all patients)
- $f(L_2|A_0, A_1 = 0, L_0, L_1)$ (only patients ICE free following visit 1)
- $f(Y|A_0, A_1 = 0, A_2 = 0, L_0, L_1, L_2)$ (only patients ICE free following visit 2)

since these are all we need for step 2.

This (non-standard) version of G-formula is more robust, but less efficient statistically than the first implementation.

G-formula - reflections

After fitting required models, G-formula discards all observed data and simulates new data for all patients.

We then analyse the simulated data.

I anticipate (legitimate) hesitancy to this - can we really base our analysis in the end on a simulated dataset? This seems crazy!

Outline

Causal inference and G-formula

Missing data approaches

Conclusions

Missing data approaches

Recall the standard approach excludes data on HbA1c after ICE occurs, and fits MMRM to repeated measures of HbA1c assuming missing values are missing at random (MAR).

If based on same data and model assumptions, multiple imputation (MI) and MMRM are (essentially) equivalent [1].

Let's consider MI, where we impute the post-ICE HbA1c values.

How does this compare to the G-formula method?

G-formula and MI

	G-formula v1	G-formula v2	MI
Data used to fit imp. models	Pre and post-ICE	Pre-ICE	Pre-ICE
Data imputed	All times for all patients	All times for all patients	Post-ICE times in patients with ICE

G-formula v1 - ‘standard’ G-formula
G-formula v2 - modified G-formula where we only fit models using ICE free patients at each visit

G-formula v2 and MI still differ - G-formula replaces all observed data with simulated/imputed values, whereas MI only imputes post ICE data.

G-formula and MI equivalence

In fact, at least for certain (important) model setups, G-formula v2 and MI are the same.

Both methods impute final outcomes for patients who experience ICE, based on same model fits.

For patients with no ICE, the mean of their imputed values in G-formula v2 matches the mean of their observed values (as used by MI).

This is basically because in regression, the mean of the fitted values equals the mean of the dependent variable in the sample.

This can also be used to argue that G-formula v2 and MMRM are the same (see Olarte Parra *et al* 2022 [4]).

G-formula via multiple imputation

Is it possible to use MI methods but exploit observed post-ICE data when fitting models, like ‘standard’ G-formula (v1)?

Yes - based on ideas from using MI to create synthetic datasets.

This can be really useful, because we can use MI to handle both missing data and missing (no ICE) potential outcomes.

We propose the G-formula via multiple imputation algorithm...

Data setup

To perform G-formula via MI, add n_{syn} new rows to the dataset, setting Ls and Y to missing, and treatment indicators to desired treatment regime.

E.g. for randomised treatment $A_0 = 1$ and no ICE subsequently:

R	L_0	A_0	L_1	A_1	L_2	A_2	Y
1	5.3	0	0.5	0	2.2	0	1.3
1	7.3	1	4.2	0	4.6	0	5.5
1	6.5	1	0.4	1	0.8	1	1.9
1	8.1	0	1.6	1	4.1	1	7.0
0	NA	1	NA	0	NA	0	NA
0	NA	1	NA	0	NA	0	NA
0	NA	1	NA	0	NA	0	NA
0	NA	1	NA	0	NA	0	NA

R denotes whether the row was originally observed (=1) or not (=0).

Imputation of potential outcomes

We then use MI software to impute missing potential outcomes in the new synthetic rows, according to our sequential models per G-formula.

Note, since the missingness pattern is monotone, iterative methods are not required.

E.g. in SAS, use monotone imputation method, or in R using `mice`, specify `maxit=1` and the required custom predictor matrix.

Analysis of imputed datasets

We then analysis Y in the synthetic rows ($R = 0$), e.g. by calculate the mean of Y, to estimate $E(Y^{1,0,0})$.

Rubin's variance estimator overestimates (substantially) the variance of this estimator.

This is because of uncongeniality - the analysis is only using a subset of the records.

Instead we must use variance estimator developed for synthetic datasets [5]:

$$\hat{V}_{\text{syn}} = (1 + M^{-1})\hat{B} - \hat{V}$$

Note: this is between minus within variance, rather than usual between plus within variance!

Missing data

What about if we have some missing data in the original dataset?

First, create M imputations of missing values in original dataset.

Then augment each imputation with n_{syn} rows as before, and impute each of these once, resulting in M imputations.

Analyse as described previously using modified variance estimator.

Comparing treatments

To compare randomised treatments, add additional synthetic rows, with A_0 set to the other treatment.

Estimate $E(Y^{0,0,0})$ using synthetic rows with $A_0 = 0$ and $E(Y^{1,0,0})$ using synthetic rows with $A_0 = 1$.

Alternatively, we can fit a regression model for Y with A_0 and baseline variables as covariates for improved precision.

G-formula via MI

Carrying out G-formula (using post-ICE data) via MI methods is attractive compared to standard G-formula implementations:

- avoids the need to use bootstrapping methods
- leverages existing understanding and software for multiple imputation
- readily accommodates imputation of missing data

Outline

Causal inference and G-formula

Missing data approaches

Conclusions

Conclusions

- Hypothetical estimands require careful specification to be well defined and relevant
- For estimation, need to adjust for all common causes of ICE and final outcome
- Post-ICE data can be exploited for improved power, but this requires more modelling assumptions
- MMRM and MI discarding post-ICE data can be viewed as particular implementations of G-formula method from causal inference
- MI can be adapted to exploit post-ICE data, providing a convenient route to handling missing actual and counterfactual data
- Further research needed to understand relationship between G-formula and G-estimation (see Florian's talk) - in some cases they turn out to be identical [6]

References I

[1] J. R. Carpenter and M. G. Kenward.
Missing data in randomised controlled trials - a practical guide.
Birmingham, UK. National Health Service Co-ordinating Centre for Research Methodology, 2007.

[2] M. A. Hernán and S. L. Taubman.
Does obesity shorten life? the importance of well-defined interventions to answer causal questions.
International journal of obesity, 32(3):S8–S14, 2008.

[3] B. Holzhauer, M. Akacha, and G. Bermann.
Choice of estimand and analysis methods in diabetes trials with rescue medication.
Pharmaceutical statistics, 14(6):433–447, 2015.

References II

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

Hypothetical estimands in clinical trials: a unification of causal inference and missing data methods.

Statistics in Biopharmaceutical Research, in press(in press):1–26, 2022.

[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] S. Vansteelandt.

Estimating direct effects in cohort and case–control studies.

Epidemiology, 20(6):851–860, 2009.