mimix: Reference-based imputation of missing data

by Ian R White, Kevin McGrath, Matteo Quartagno, Suzie M Cro and James Carpenter

Abstract Reference-based imputation is a multiple imputation technique which imputes quantitative outcome data that are missing after participant discontinuation of allocated treatment in a randomised trial. We present and describe an R package, mimix, for performing reference-based imputation, including a causal model variant, and we compare its implementation with that in SAS and Stata.

Introduction

Missing data are a challenge for many analyses. This article tackles the specific issue of a randomised trial with a repeatedly measured quantitative outcome, where participants who discontinue their randomised treatment are not followed up thereafter and hence have missing outcome data. We assume that the aim is to estimate the effect of treatment on the actual outcomes of the participants, whether observed or not: this has been termed a "de facto" estimand (Carpenter et al., 2013) or a "treatment policy" estimand (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 2019). In this setting, an analysis under the commonly used missing at random (MAR) assumption would assume that the missing (post-treatment) outcomes are comparable (conditional on observed data) with the observed (on-treatment) outcomes, and would therefore estimate the effect of treatment on the outcomes of the participants if treatment was never discontinued: this has been termed a "de jure" estimand (Carpenter et al., 2013) or a "hypothetical" estimand (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 2019).

Carpenter et al. (2013) proposed that missing data after discontinuation of allocated treatment could be imputed by assuming that post-discontinuation outcomes behave, in some sense, like the outcomes in a reference group. For example, if participants who have discontinued their allocated treatment are likely to be receiving similar treatment to participants who were allocated to control treatment, then the control group would be the reference group. Carpenter et al. (2013) proposed imputing the missing data from a joint multivariate Normal (MVN) distribution for the complete (observed and unobserved) data, and proposed five ways to construct this joint distribution: "jump to reference" (J2R), "copy reference" (CR), "copy increments in reference" (CIR), "last mean carried forward" (LMCF) and "missing at random" (MAR). All approaches start by fitting a MVN distribution to the data from each arm. As an example, the J2R joint distribution for a participant in a specific arm takes the means for that arm up to the point of treatment discontinuation and the means for the reference arm afterwards.

The implicit assumptions behind this approach were explored by White et al. (2020), who proposed a causal model in which the treatment effect after discontinuation is a specified multiple of the treatment effect at the point of discontinuation. They showed that J2R, CR and CIR are the special cases of this causal model in which the treatment effect disappears, decays or is maintained after treatment discontinuation, while LMCF and MAR are not special cases of the causal model.

The RBI methods of Carpenter et al. (2013) were implemented in SAS in what have become known as "the five macros" and are available on the web page (on www.missingdata.org.uk) of the DIA working group for missing data. These macros also provide for "delta-adjustment" in which imputed values are modified by a user-specified amount (Ratitch et al., 2013): this is useful is performing sensitivity analysis, since the RBI methods make a number of untestable assumptions (Leacy et al., 2017). The methods were then implemented in Stata by Cro et al. (2016).

We have implemented the RBI methods in a new R package, which includes the a full implementation of the causal model. The causal model was previously specified for a two-arm trial and has been extended here for a multi-arm trial. The aim of this paper is to describe our implementation of mimix in R, to illustrate its use, and to describe its novel features by comparison with the implementations in SAS and Stata.

Reference-based imputation

Setting

[nicked from White et al. (2020)] We assume that quantitative outcome measurements are scheduled at baseline and at T occasions after randomisation. Let Z be the random variable for the participant's randomised treatment arm, let $Z = z$ denote an arm in which we want to impute missing values, and $Z = r$ denote the reference arm. Let Y_t be the random variable for the participant's outcome at visit $t = 0, ..., T$. It is convenient to define $\boldsymbol{Y}_{\leq t} = (Y_0, \ldots, Y_t)$, the vector of all outcomes up to and including visit t; $\boldsymbol{Y}_{>t} = (Y_{t+1}, \ldots, Y_T)$, the vector of all outcomes after visit t; and $\boldsymbol{Y} = (Y_0, \ldots, Y_T)$, the vector of all outcomes. Let D be the random variable for the participant's last visit prior to discontinuing treatment, so $D = 0, ..., T$. Y_t is observable for all t but only observed for $t \leq D$, because we assume no off-treatment data. We aim to impute the unobserved values of Y_t for $t > D$: we stress that these are the outcomes that existed but were unobserved, not the outcomes that would have existed if treatment had been continued.

Reference-based imputation

Carpenter et al. (2013) proposed a generic MI algorithm for this setting:

1. For each treatment arm z, fit a multivariate normal model to all observed data, using a Bayesian approach with an improper prior and assuming MAR. The model has unstructured mean $\boldsymbol{\mu}_z$ and unstructured variance-covariance matrix $\boldsymbol{\Sigma}_z$.
2. For each treatment arm z, draw a mean vector $\boldsymbol{\mu}_z^*$ and variance-covariance matrix from the posterior distribution $\boldsymbol{\Sigma}_z^*$.
3. For each treatment arm z and each possible treatment discontinuation visit t, use the drawn values to build a hypothetical joint distribution of the outcomes $\boldsymbol{Y}_{\leq t}$ up to time t and the outcomes $\boldsymbol{Y}_{>t}$ after time t, using one of the methods described below. Thus a MVN distribution is built for $\boldsymbol{Y}|Z = z, D = t$. Five methods are mainly distinguished by their choice of mean:
   (a) Jump to reference (J2R): mean = $\left(\boldsymbol{\mu}^*_{z,\leq t}, \boldsymbol{\mu}^*_{r,>t}\right)$.
   (b) Copy reference (CR): mean = $\boldsymbol{\mu}^*_r$.
   (c) Copy increments in reference (CIR): mean = $\left(\boldsymbol{\mu}^*_{z,\leq t}, \{\mu^*_{z,t} - \mu^*_{r,t}\}\boldsymbol{e}_{T-t} + \boldsymbol{\mu}^*_{r,>t}\right)$ where $\boldsymbol{e}_p$ is a row vector $(1, \ldots, 1)$ of length p.
   (d) Missing at random (MAR): mean = $\boldsymbol{\mu}^*_z$.
   (e) Last mean carried forward (LMCF): mean = $\left(\boldsymbol{\mu}^*_{z,\leq t}, \mu^*_{z,t}\boldsymbol{e}_{T-t}\right)$.

   CRK proposed corresponding variance matrices. Following White et al. (2020), we describe only the regression coefficient matrix and conditional variance matrix of the potential outcomes after visit t given those before visit t. CRK set these to be taken from arm z for MAR and LMCF, and from arm r for J2R, CIR and CR.

   [KEVIN: An approach that we call *RBI alternative* instead uses $\boldsymbol{\beta}_t(T)$ and $\boldsymbol{\Omega}_t(T)$ for all RBI methods: did we code this???]
4. For each treatment arm z and each observed treatment discontinuation visit t, construct the imputation distribution of $\boldsymbol{Y}_{>t}$ given $\boldsymbol{Y}_{\leq t}$. Sample $\boldsymbol{Y}_{>t}$ from this conditional distribution, to create a "completed" data set.
5. Repeat steps 2-4 m times, resulting in m imputed data sets.
6. Analyse each imputed data set and combine the results using Rubin's rules (Rubin, 1987).

Causal model

The causal model has previously been stated for two arms only. Here we extend its statement to the multi-arm case. We define the potential outcome $Y_t(z, s)$ at visit t as the outcome that would have been observable if, possibly contrary to fact, a participant received active treatment z for s periods only. In particular, $Y_t(z, 0)$ is the potential outcome if never treated: it is the same for all z and is written $Y_t(0)$. Similarly $Y_t(z, T)$ is the potential outcome if always treated with z. We define $\boldsymbol{Y}_{\leq t}(z, s)$, $\boldsymbol{Y}_{>t}(z, s)$ and $\boldsymbol{Y}(z, s)$ as before. We let $\mu_t(z, s) = \mathbb{E}[Y_t(z, s)]$, the mean of the potential outcome at visit t if active treatment z were received for s periods only. Similarly we define $\boldsymbol{\mu}_{\leq t}(z, s)$, $\boldsymbol{\mu}_{>t}(z, s)$ and $\boldsymbol{\mu}(z, s)$.

[Omit?:] The variance-covariance matrix of the potential outcomes is $\boldsymbol{\Sigma}(s) = \text{var}\left(\boldsymbol{Y}(s)\right)$ with sub-matrices $\boldsymbol{\Sigma}_{\leq t \leq t}(s)$, $\boldsymbol{\Sigma}_{>t>t}(s)$ and $\boldsymbol{\Sigma}_{>t\leq t}(s)$. We define the matrix of regression coefficients of potential outcomes after visit t on those up to visit t as $\boldsymbol{\beta}_t(s) = \boldsymbol{\Sigma}_{>t\leq t}(s)\boldsymbol{\Sigma}_{\leq t\leq t}(s)^{-1}$, and the residual variance of the potential outcomes after visit t given those up to visit t as $\boldsymbol{\Omega}_t(s) = \boldsymbol{\Sigma}_{>t\leq t}(s)\boldsymbol{\Sigma}_{\leq t\leq t}(s)^{-1}\boldsymbol{\Sigma}_{>t\leq t}(s)^T$.

The key model assumption describes how the maintained effect of treatment after discontinuation relates to the effect of treatment before discontinuation:

$$\boldsymbol{\mu}_{>t}(z,t) - \boldsymbol{\mu}_{>t}(0) = \boldsymbol{K}_{z,t}\left\{\boldsymbol{\mu}_{\leq t}(z,t) - \boldsymbol{\mu}_{\leq t}(0)\right\} \tag{1}$$

where $\boldsymbol{K}_{z,t}$ is a $(T-t)\times(t+1)$ matrix of sensitivity parameters: it is not identified by the data and must be specified by the user. In practice we make a choice of $\boldsymbol{K}_{z,t}$ determined by just two parameters (k_0, k_1) giving the simplified model for each $u > t$:

$$\boldsymbol{\mu}_u(z,t) - \boldsymbol{\mu}_u(0) \quad = \quad k_0 k_1^{v_u - v_t}\left\{\boldsymbol{\mu}_t(z,t) - \boldsymbol{\mu}_t(0)\right\} \tag{2}$$

where v_t is the time (on a suitable scale) of visit t.

[KEVIN - can we let (k_0, k_1) depend on arm?]

Estimation involves other assumptions which make explicit the ideas of Carpenter et al. (2013): that randomisation is independent of all potential outcomes; that step 1 of the RBI algorithm in the active arms estimates the distribution of $\boldsymbol{Y}(z,s)$, and in the reference arm estimates the distribution of $\boldsymbol{Y}(0)$; that the conditional distributions follow linear regressions (which is true if the joint distribution is MVN); and that treatment discontinuation is unaffected by future potential partly-treated outcomes.

These give the mean of the imputation model:

$$\begin{aligned} &E[\boldsymbol{Y}_{>t}(z,t)|Z = z, \boldsymbol{Y}_{\leq t}, D = t] \\ &= \boldsymbol{\beta}_t(z,t)\left\{\boldsymbol{Y}_{\leq t} - \boldsymbol{\mu}_{\leq t}(z,t)\right\} + \boldsymbol{K}_{z,t}\left\{\boldsymbol{\mu}_{\leq t}(z,t) - \boldsymbol{\mu}_{\leq t}(0)\right\} + \boldsymbol{\mu}_{>t}(0). \end{aligned} \tag{3}$$

[Again we can handle variances... how much to say?]

Delta adjustment

We follow the approach of the five macros. Any imputed value $Y^*_{z,t}$ is replaced by $Y^*_{z,t} + \sum_{u=t+1}^{u=T} a_u b_{u-t}$ where a_u is a user-specified shift for time u and b_w is a user-specified scaling multiplier that controls how the user-defined shift for time u is applied to all times $\geq u$. For example, if $b = (1, 0.7, 0.5, ...)$ then a participant who discontinues treatment just after visit t has their imputed value at time $t+1$ incremented by a_{t+1}, their imputed value at time $t+2$ incremented by $0.7a_{t+1} + a_{t+2}$, their imputed value at time $t+3$ incremented by $0.5a_{t+1} + 0.7a_{t+2} + a_{t+3}$, and so on. Delta adjustment can apply either with RBI or the causal method and is always applied after imputation.

Interim missing values and covariates

[Kevin - to discuss]

Package

KEVIN TO DRAFT - INITIALLY TAKE FROM HELP.

Arguments

Outputs

Examples

KEVIN TO DRAFT. Which data to use?

Comparisons with other packages

KEVIN TO DRAFT – THESE ARE JUST MY NOTES. Would a table be suitable?

1. flavours of LMCF in 5 macros
2. handling participants with no observed outcomes
3. how baseline covariates are modelled: SAS, can interact with time; Stata, R, can't
4. handling of interim missing values
5. anything to say about algorithm differences? prior choice?

Limitations and discussion

- flexible package
- doesn't cover other outcome types; e.g. methods have been proposed for recurrent events (Keene et al., 2014)
- trials should be designed consistently with their estimand, so if the treatment-policy estimand is of interest then outcomes should be collected after treatment discontinuation. Analysis options for this setting are still in development: options include including treatment discontinuation time in the model and imputing under MAR [ref Tom at GSK]; using RBI or causal model and reserving the post-discontinuation data for model checking; or using RBI or causal model and using the post-discontinuation data to estimate model parameters such as $K_{z,t}$.

Bibliography

J. R. Carpenter, J. H. Roger, and M. G. Kenward. Analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions, and inference via multiple imputation. *Journal of Biopharmaceutical Statistics*, 23(3):1352–71, 2013. ISSN 1520-5711. doi: 10.1080/10543406.2013.834911. URL http://www.tandfonline.com/doi/abs/10.1080/10543406.2013.834911. [p1, 2, 3]

S. Cro, T. P. Morris, M. G. Kenward, and J. R. Carpenter. Reference-based sensitivity analysis via multiple imputation for longitudinal trials with protocol deviation. *Stata Journal*, 16(2):443–463, 2016. [p1]

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials E9(R1). Technical report, 2019. URL https://www.ich.org/page/efficacy-guidelines{#}9-2. [p1]

O. N. Keene, J. H. Roger, B. F. Hartley, and M. G. Kenward. Missing data sensitivity analysis for recurrent event data using controlled imputation. *Pharmaceutical Statistics*, 13(4):258–264, jul 2014. ISSN 15391604. doi: 10.1002/pst.1624. URL http://doi.wiley.com/10.1002/pst.1624. [p4]

F. P. Leacy, S. Floyd, T. A. Yates, and I. R. White. Analyses of Sensitivity to the Missing-at-Random Assumption Using Multiple Imputation With Delta Adjustment: Application to a Tuberculosis/HIV Prevalence Survey With Incomplete HIV-Status Data. *American Journal of Epidemiology*, 185(4):304–315, jan 2017. ISSN 0002-9262. doi: 10.1093/aje/kww107. URL https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kww107http://aje.oxfordjournals.org/lookup/doi/10.1093/aje/kww107. [p1]

B. Ratitch, M. O'Kelly, and R. Tosiello. Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. *Pharmaceutical Statistics*, 12:337–347, 2013. URL http://onlinelibrary.wiley.com/doi/10.1002/pst.1549/abstract. [p1]

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, New York, 1987. [p2]

I. R. White, R. Joseph, and N. Best. A causal modelling framework for reference-based imputation and tipping point analysis in clinical trials with quantitative outcome. *Journal of Biopharmaceutical Statistics*, 30(2):334–350, mar 2020. ISSN 1054-3406. doi: 10.1080/10543406.2019.1684308. URL https://www.tandfonline.com/doi/full/10.1080/10543406.2019.1684308. [p1, 2]

Ian R White
MRC Clinical Trials Unit at UCL
90 High Holborn, 2nd Floor, London WC1V 6LJ
UK

ORCiD: 0000-0002-6718-7661
ian.white@ucl.ac.uk

Kevin McGrath
MRC Clinical Trials Unit at UCL
90 High Holborn, 2nd Floor, London WC1V 6LJ
UK
(ORCiD if desired)
kevin.mcgrath@ucl.ac.uk

Matteo Quartagno
MRC Clinical Trials Unit at UCL
90 High Holborn, 2nd Floor, London WC1V 6LJ
UK
(ORCiD if desired)
m.quartagno@ucl.ac.uk

Suzie M Cro
Imperial Clinical Trials Unit
Imperial College London, 1st Floor, Stadium House London, W12 7RH
UK
(ORCiD if desired)
s.cro@imperial.ac.uk

James Carpenter
MRC Clinical Trials Unit at UCL
90 High Holborn, 2nd Floor, London WC1V 6LJ
UK
(ORCiD if desired)
j.carpenter@ucl.ac.uk