Are Missing Data Properly Accounted for in Health Economics and Outcomes Research?

Gianluca Baio, PhD, MSc, University College London, London, England, UK; **Necdet B. Gunsoy, PhD, MPH**, Global Market Access and Pricing, AbbVie, Maidenhead, Berkshire, UK; **Nneka Onwudiwe, MBA, PharmD, PhD**, US Food and Drug Administration, Silver Spring, MD, USA; and **David J. Vanness, PhD**, The Pennsylvania State University, State College, PA, USA on behalf of the ISPOR Missing Data in HEOR Working Group of the ISPOR Statistical Methods Special Interest Group

The ISPOR Statistical Methods in Health Economics and Outcomes Research (HEOR) Special Interest Group investigated how missing data impact the analysis of HEOR data, and the potential of using full Bayesian methods to account for partially recorded information in the broader context of economic modeling.

When critical data go missing

When the ISPOR Statistical Methods in Health Economics and Outcomes Research (HEOR) Special Interest Group (SIG) convened to choose its first key project, members of the leadership group expressed interest in developing guidance for a variety of statistical approaches. One theme emerged as both central to, and uniquely challenging in, our field: how to deal with missing data, particularly in the context of real-world evidence.

Much of the field of HEOR boils down to the process of making statistical inference from clinical and economic data so that decision makers can assess the value of medical care. The rules for making valid inference are challenging enough in the world of clinical trials, where outcomes are assessed under carefully controlled conditions in idealized settings designed to capture all necessary data. While such randomized controlled trial data are used in HEOR, we also often contend with complex patterns of treatment delivery, each treatment having potentially heterogeneous effects, and resulting in multiple causally interconnected outcomes. Many of our outcomes are subjective, noisy, and logistically or economically burdensome to collect.

Many of our data sources were designed to facilitate clinical care, or generate bills conforming to administrative rules and not for conducting scientific research. The very nature of data used in our field both makes missing data more likely, and amplifies its potential to create bias.

Rubin provides a useful framework for understanding missingness.[1] Imagine that there's a process that gives rise to the complete, idealized, true dataset—everything we need to conduct valid inference for our purpose. Now, imagine a second process that causes some of those data to be deleted. It might be entire variables that go missing for everyone, or they might be missing for some individuals, but not others. Some individuals might be missing all their data and it's as if they never existed, while others are missing only some variables. In the context of longitudinal data, such missingness may come and go, or it may persist. The patterns themselves are interesting, but it is how the data went missing in the first place that matters most.

When the process that caused the data to go missing is unrelated to the process that generated the true dataset, we say that data is "missing completely at random." In this case, ignoring observations with missing data reduces our inferential power, but it does not create bias; the estimate effect remains unbiased but its precision decreases (standard error increases). We can make up for it by just collecting more of the same data. Sometimes, the process that caused the data to go missing is related to the process that generated the true dataset, but in a way we can control for by using data we actually observed. We term this process "missing at random;" the observed data are reweighted and missing values are imputed based on their relationship to the observed data.[2] Alternatively, we may attempt to model the 2 processes explicitly. But when the process that caused some of the data to go missing depends on the missing data themselves, things get especially challenging. In this case, we say that the underlying generating process is "missing

> ...a full Bayesian model accounting for missing data extends the industry standard tool of "multiple imputation," where missing values are replaced by simulations obtained from the whole system of modelling assumptions.

not at random," and we need to appeal to "extra-statistical" assumptions about the missing data process in order to improve our chances of getting our inference correct.

Our key project is reviewing and synthesizing the methods literature to see what approaches have already been used and assessed in HEOR and, in addition, what we can adapt from other fields to suit our unique challenges in HEOR.

Preliminary literature review

A literature review was conducted to understand methodological approaches used to account for missing data in cost-effectiveness analyses specifically. Studies were eligible only if they focused on addressing missing data for costs, health-related quality of life (HRQoL) measures, or utility.

We conducted the search in PubMed. Identified records were screened independently by authors Gunsoy and Baio. Data from eligible abstracts were then extracted by volunteers from the SIG membership. A total of 16 studies were identified for which information on the context, missing data, and method(s) used were extracted.

Identified studies were mainly conducted in clinical trials, addressing either missing cost, HRQoL, or both jointly, and evaluated multiple imputation. The majority of studies were focused on imputing outcomes rather than explanatory factors. Although most studies applied multiple imputation, a large variety of model specifications were observed.

Practical guidance on how to handle missing data

Table 1 presents a list of recommendations from 3 published studies identified in the literature review that address missing data within the context of a clinical trial. The 3 studies were chosen because of their relevance to handling missing data in the field of HEOR and for decision makers tasked with comparing and choosing among available treatment options. As a practical guide, the studies presented recommend that sensitivity analysis be performed to determine to what extent trial results change to different missing data assumptions. The cost-effectiveness analysis-based studies in Table 1 recommend 2 model-based "missing not at random" methods to handling sensitivity analysis, the selection models and pattern mixture models approach.

Next steps

As suggested by recent reviews,[3,4] over the past few years the HEOR literature seems to have caught up with other research fields in understanding the importance of correct reporting, analysis, and statement of limitations when it comes to recommendations for decision making based on data affected by missingness.

However, HEOR data are characterized by extra complexity, including a bivariate outcome (clinical benefits and costs), whose components are likely to be correlated and characterized by asymmetric distributions and spikes (eg, excess of 0 costs or 1 utilities). Failure to properly account for these elements may produce output from the underlying statistical model that possibly underestimates the underlying uncertainty in the actual cost-effectiveness profile of a given intervention. This in turn has the potential to make the whole decision-making process flawed, as it is clearly based on untenable premises.

The way forward is to embrace this complexity and use statistical methods that are suitable to deal with the many nuisances of the data we analyze. In particular, Bayesian methods are increasingly popular in HEOR,[5-7] (eg, when dealing with network meta-analysis and evidence synthesis[9], analysis of survival data[10] and decision making in general[8]) and are a promising tool to deal with missing values.

In a nutshell, the fundamental feature of a Bayesian analysis is that uncertainty is modeled using the language of probability distributions. Much as in a standard, "frequentist" analysis, sampling variability surrounding the observed data is modelled using a distribution (eg, a Beta distribution to model quality-adjusted life years [QALYs], or a Gamma distribution for the observed costs). However, a Bayesian analysis implies a probability distribution for any quantity that is not deterministically known. This includes: (a) *model parameters* (eg, means, population incidence, etc) that we shall never be in a position of observing directly; and (b) as *yet unobserved data* (eg, that can be obtained using real-world evidence produced by registries of clinical practice). These may or may not be available in the future and thus we are still uncertain about what their value will be.

In a Bayesian sense, the missing data process is simply another part of a wider model, which considers the 2 outcomes of interest, as well as any other relevant covariate. The objective of the analysis is to specify a joint probability distribution for the (partially) observed data (including benefits, costs, and a missingness indicator) and the model parameters, which typically indicate the population average costs and benefits. Modeling assumptions are made explicit in terms of *prior* probability distributions describing possibly subjective

Our key project is reviewing and synthesizing the methods literature to see what approaches have already been used and assessed in HEOR and, in addition, what we can adapt from other fields to suit our unique challenges in HEOR.

knowledge on the model parameters, as well as probability distribution to describe variability in the (partially) observed data. Combining these with the evidence provided by the data, we can revise our assessment of the uncertainty underlying the unobserved quantities in the model (eg, the population average costs and benefits). The updated, *posterior* distribution can then be used directly to aid the decision-making process.

Of MICE and missing data

In effect, a full Bayesian model accounting for missing data extends the industry standard tool of "multiple imputation," where missing values are replaced by simulations obtained from the whole system of modelling assumptions. In fact, Rubin's original ideas were arguably very Bayesian in nature, but

Table 1. Handling Missing Data Within the Context of a Clinical Trial

Publication Year	Study Objective	Outcome Measure(s)	Recommendations
2014	A review article (Faria et al) that provides guidance on how to handle missing data in within-trial CEAs following (i) a plausible assumption for the missing data mechanism; (ii) the method chosen for the base-case; and (iii) sensitivity analysis	CEA	**Stage 1:** Descriptive Analysis of Missing Data Mechanism **Descriptive analysis of the missing data** 1) Amount of missing data by trial group at each follow-up period 2) Missing data patterns 3) Association between missingness and baseline variables 4) Association between missingness and observed outcomes **Stage 2:** Choosing Between Alternative Methods Given Their Underlying Assumptions **Handling Missing Baseline Values** 1) Mean imputation and MI are suggested options **Complete Case Analysis, Available Case Analysis, and Inverse Probability Weighting** 1) CCA and available case analyses are valid under MCAR 2) CCA is a good starting point and benchmark but not for the base case 3) Available case analysis makes more efficient use of the data compared to CCA 4) IPW is suitable for a monotonic pattern of missing data **Single Imputation** 1) Mean imputation valid for missing baseline variables 2) Conditional regression imputation assumes MAR but can affect the cost-effectiveness estimate 3) Last-value carried forward (LVCF) can bias parameter estimates 4) Single imputation methods are not appropriate to handle missing data on outcomes **Multiple Imputation** 1) MI can handle both monotonic and nonmonotonic missing data under MAR and can be modified to handle MNAR 2) Two approaches to implementing MI: joint modelling (MI-JM) and chained equations (MICE) 3) MI-JM assumes multivariate normal distribution 4) MICE accommodates non-normal distributions, allows for interactions and nonlinear terms, and incorporates variables that are functions of imputed variables 5) MICE can handle datasets with a large number of variables with missing data 6) MICE is more applicable to missing data in within-trial CEAs **Likelihood-Based Methods** 1) Likelihood-based models assume MAR conditional on the variables, unless MNAR is explicitly modeled 2) Likelihood-based methods can produce similar results to MI when all variables that relate to missingness are included in the analysis model 3) Relies on the correct specification of the model; the impact of different specifications should be compared and reported **Stage 3:** Methods for Sensitivity Analysis to MAR Assumption 1) Selection models and pattern mixture approaches 2) Selection models using a weighting approach tends to fail for large departures from MAR
2018	A review article (Leurent et al.) to determine the extent of missing data, how they were addressed in the analysis, and whether sensitivity analyses to different missing data assumptions were performed in studies identified. Also, to provide recommendations to improve practice.	CEA	**Prevent** 1) Maximize response rate (consider questionnaire design, mode of administration, reminders, incentives, participants' engagement, etc.) 2) Consider alternative data sources (eg, routinely collected data) 3) Monitor cost-effectiveness data completeness while trial ongoing **Primary** 1) Formulate realistic and accessible missing data assumption for the primary analysis (typically, but not necessarily, a form of the missing at random assumption) 2) Use appropriate method valid under that assumption (typically, but not necessarily, multiple imputation or maximum likelihood) **Sensitivity** 1) Discuss with clinicians and investigators to formulate plausible departures from the primary missing data assumption 2) Consider a broad range of assumptions, including missing not a random 3) Use appropriate method valid under these assumptions (typically, but not necessarily, pattern-mixture models or a reference-based approach) **Report** 1) Report the number of participants with cost and outcome data, by arm and time-point 2) Report possible reasons for nonresponse and baseline predictors of missing values 3) Describe methods used, and underlying missing data assumptions 4) Draw overall concusion in light of the different resuts and the plausabiity of the respective assumptions
2018	A review article (Rombach et al.) that provides guidance on the choice of MI models for handling missing PROMs data based on the characteristics of the trial dataset, specifically with regards to the use of MI.	PROMs	1) Imputation at the item level may not be feasible for small sample sizes and/or larger proportions of missing data 2) Smaller samples with large amounts of missing data, imputation at the composite score level is more beneficial when there is a predominantly unit-nonresponse pattern 3) When performing imputation at the item level using ordinal logit models, the dataset should be investigated thoroughly for low count and potential problems due to perfect prediction 4) Ideally, imputation at the item/subscale level may provide more precise estimates of treatment effect compared to the imputation at the composite score level or CCA but it's often unfeasible and prone to convergence

Abbreviations: CCA indicates complete cases analysis; CEA, cost-effectiveness analysis; IPW, inverse probability weighting; MAR, missing at random; MI, multiple imputation; MICE, multiple imputation by chained equations; ML, maximum likelihood; MNAR, missing not at random; PROMs, patient-reported outcome measures.

at the time, there simply wasn't the computer power and methodology to perform the computations.[1] Thus, commonly used methods (eg, multiple imputation by chained equation, [MICE])[11] are based on a hybrid of Bayesian grounding and frequentist implementation. Crucially, these methods are often devised for modelling structures that are slightly simpler than those we need to face in HEOR (eg, when the interest is only in a single outcome variable or when the data are more well-behaved and can be reasonably modeled using normal distributions). For this reason, expanding them to a full Bayesian approach may be a very attractive way forward for our field. This, coupled with the increasing drive to using suitable statistical software and appropriately sophisticated models throughout the statistical and economic analysis, indeed has the potential to improve the decision-making process.

Disclaimer: This article reflects the views of the author and should not be construed to represent FDA's views or policies.

References

1. Rubin DB. Inference and missing data. *Biometrika*. 1976;6(3):581–592. https://doi.org/10.1093/biomet/63.3.581

2. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: John Wiley & Sons; 1987.

3. Faria R, Gomes M, Epstein D, White IR. A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials. *Pharmacoeconomics*. 2014;32(12):1157-1170.

4. Leurent B, Gomes M, Carpenter JR. Missing data in trial-based cost-effectiveness analysis: an incomplete journey. *Health Econ*. 2018;27(6):1024-1040.

5. Rombach I, Gray AM, Jenkinson C, Murray DW, Rivero-Arias O. Multiple imputation for patient reported outcome measures in randomised controlled trials: advantages and disadvantages of imputing at the item, subscale or composite score level. *BMC Med Res Methodol*. 2018;18(1):87-103.

6. Gabrio A, Mason A, Baio G. A full Bayesian model to handle structural ones and missingness in economic evaluations from individual-level data. *Stat Med*. 2019;38(8):1-22.

7. Spiegelhalter D, Abrams K, Myles JP. *Bayesian Approaches to Clinical Trials and Healthcare Evaluation*. Hoboken, NJ: John Wiley & Sons; 2004.

8. Welton NJ, Sutton AJ, Cooper NJ, Abrams KR, Ades AE. *Evidence Synthesis for Decision Making in Healthcare*. West Sussex, United Kingdom: John Wiley & Sons. 2012.

9. Baio G. survHE: Survival analysis for health economic evaluation and cost-effectiveness analysis. J. Stat. Software (in press; available at: https://github.com/giabaio/survHE/blob/master/inst/doc/survHE.pdf); 2020.

10. Baio G. *Bayesian Methods in Health Economics*. London, England: Chapman & Hall/CRC Biostatistics Series; 2012.

11. Van Buuren, S. *Flexible Imputation of Missing Data*. Second ed. London, England: Chapman & Hall/CRC; 2018.

Additional information

This article was developed as a collaborative effort of members of the ISPOR Statistical Methods in Health Economics and Outcomes Research Special Interest Group (SIG). We thank all reviewers who commented during our ISPOR conference presentation at ISPOR Europe 2018. We especially thank the individuals who submitted written comments on drafts of this article. The following leadership members of the SIG submitted feedback (in alphabetical order): Bijan Borah, Joseph Cappelleri, Christoph Gerlinger, Helene Karcher, Rita Kristy, Jennifer Lord-Bessen, Sri Ram Pentakota, and Jessica Roydhouse. Source of financial support: None of the authors received support for their participation in this SIG. All authors volunteered their time for discussion, research, and writing of this article. This research was supported in part by ISPOR.

To view the presentation, go to https://www.ispor.org/docs/default-source/presentations/91681pdf.pdf?sfvrsn=ce71ee4d_0.