York Health Economics Consortium

Recommended standards for managing and reporting missing utility data for health technology appraisal

Supplementary Material

NEIL HANSELL, Statistician

KARIN BUTLER, Head of Technical Research

Dr Joe Moss, Principal Statistician

18/11/2024

Table of Contents

1 Supplementary Material A – Missingness Mechanisms Definitions

Missingness mechanism defined as described by Rubin (1976) [1].

1.1 Missing completely at random (MCAR)

Missingness based on no observed or unobserved data characteristics. In this setting, there's no underlying relationship between missingness utility and any observed or unobserved data sets. Missing data is completely random.

1.2 Missing at random (MAR)

Missingness caused by observed information within the data [2]. In this setting, MAR data was generated such that those who had disease progression were more likely to be missing. Disease progression was a significant and substantial predictor of utility. Thus, MAR data was designed such that a strong predictor of utility was the cause of missingness in this setting. This simulates a plausible scenario in which those who have progressed disease are potentially less motivated, or perhaps physically unable to continue with follow-up appointments, hence they have missing data.

1.3 Missing not at random (MNAR)

Missingness caused by the unobserved values of the outcome itself [2]. In this setting, MNAR data was generated such that those who had low utility were more likely to be missing. Thus, MNAR data was designed such those in poor health are more likely to be missing. This simulates a plausible scenario in which those who have poor health are, like MAR, potentially less motivated, or perhaps physically unable to continue with follow-up appointments, hence they have missing data. However, the key distinction here is that MNAR is based on the unobserved values of the outcome, whilst MAR can be attributed to a key predictor of utility. Therefore, methods used to deal with missing data in these settings will generate different results.

2 Supplementary Material B – Analytical Methods Explanations

2.1 Complete case analysis (CCA)

A CCA simply involves using only those with complete data to conduct analyses. It tends to be quite a popular method to conduct analyses, likely due to it being simple and involves only using real study data with no imputation required. In this case, only cases with complete utility data were used to estimate true utility for each level and type of missingness.

2.2 Mean score estimation (MSE)

MSE is an imputation method that involves imputing the mean for missing data. The mean can either be imputed for the overall observed data, or by certain groups. In this setting, disease progression, by design was a significant predictor of utility, thus MSE was conducted based on progression status. Mean utility was calculated based on the observed data for those who were and weren't progressed and used to impute missing utility data.

Despite us having the knowledge of the strong relationship between disease progression and utility as we created the simulated dataset, it would be reasonable for an analyst to examine this dataset and recognise this strong relationship, thus, MSE stratified by disease progression status would be a justifiable approach in this case.

2.3 Multiple imputation via chained equations (MICE)

Whilst MSE is a singular imputation method (i.e. we have one imputed value for each missing data point), MICE is an analytical technique that estimates missing data by creating multiple plausible values for each missing data point using regression equations [3]. In our case, this was done 5 times to create 5 complete datasets. A maximum of 20 iterations were conducted for each missing value within each complete dataset to refine the imputed values and account for uncertainty.

MICE is the framework in which the imputation method occurs. The method used is predictive mean matching (PMM). PMM uses regression equations for the outcome (utility) against key predictors so that, for each of the 5 imputations and 20 iterations, utility is imputed by matching individuals with missing utility to individuals with observed utility. For each missing value, PMM identifies an arbitrary number of individuals with similar values for the predictors. Then it matches one of these individuals at random, so that the imputed value is an observed value of the outcome where predictions are closely matched [3].

2.4 Linear mixed model (LMM)

A generalised linear mixed model (GLMM) with a beta distribution and a logit link were used to estimate utility based on key predictors. The same predictors that were used to generate utility scores in the simulated data were also used here in the GLMM. Unique patient IDs were used as random effects within the GLMM to account for repeated measures for each individual (multiple follow-ups).

3 References

1. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581-92.
2. Ramzi W. Nahhas. Introduction to Regression Methods for Public Health Using R. 2024. Available from: https://bookdown.org/rwnahhas/RMPH/mi-mechanisms.html.
3. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine. 2011;30(4):377-99.