KING'S College LONDON NIHR Maudsley Biomedical Research Centre

Handling missing data for clinical prediction models. Review of existing methods

Internal report for NIHR Maudsley Biomedical Research Centre.

Diana Shamsutdinova[1*], Harper Clees-Baron[1], Xiniy Zhou[1], Daniel Stahl[1]

[1] Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

* diana.2.shamsutdinova@kcl.ac.uk

1 Missing Data Concepts

Most of the clinical data have missing values. This can be due to participants missing study appointments, manual or technical mistakes while filling in electronical health records, or issues during data extraction and pre-processing such as merging incompatible data sources, and coding errors.

As defined by the pioneer of the missing data methods, Donald Rubin, the following three missingness mechanisms can be recognised(1):

- **Missing completely at random (MCAR)** is a scenario when missingness is independent from any of participants' characteristics. For example, blood tests that are damaged by a laboratory.
- A more general scenario is **missing at random (MAR)**, in which missingness is assumed to be predicted by other observed data. For example, people with high income may be more likely to refuse to disclose income, while, once adjusted for another socio-economic status variable such as area-based deprivation index and education, the missingness is random.
- Finally, **missing not at random (MNAR)** is when missingness depends on an unobserved participant's characteristic or outcome. One of the MNAS scenarios is informative missingness, in which missing data is related to what is being omitted, such as participants avoiding a drug test when suspecting positive result. MNAR is the most difficult case to handle, and Buurnen advises trying to

reduce it to MAR by finding a related cause, or test results robustness by running sensitivity analyses (2).

Following (3), let R_{ij} be the missingness indicator for j^{th} feature of an i^{th} subject, x_{ij}, which is 1 if x_{ij} is missing and 0 if observed; R_i is the vector of missingness for i^{th} subject, $R_i = (R_{i1}, \ldots, R_{ik})$; subset of observed features is x_{i_obs}, and missing x_{i_m}. In context of missing values, "features" include all variables intended to be collected for an individual, which also includes outcomes. Then,

$$MCAR: P(R_i|x_i) = P(R_i)$$

$$MAR: P(R_i|x_i) = P(R_i|x_{i\_obs})$$

$$MNAR: P(R_i|x_i) \neq P(R_i|x_{i\_obs}) \qquad (2.77)$$

The simplest approach to missing data could be analysing **complete cases** only. However, this approach undermines sample's statistical power, impairs sample representativeness and may produce biased results in all but MCAR scenarios, which may not be easy to justify (4).

2 Single imputation methods

Another way to deal with incomplete data is **imputation** (filling in), which is replacing the missing data with its plausible values prior to the analyses. Mean and medium imputations can be used, though this may also lead to biased results and undervalued confidence intervals (5). Moreover, mean imputation disregards information that other features may contain about this variable. For example, weight may be better predicted while accounting for sex and height of a participant. To transfer this information into imputations, one could use any regression or classification model, including ML, and impute missing data by treating a feature with missing data as an outcome, and other features as predictors, that is, perform a regression imputation, or **"conditional mean imputation"**.

K-nearest neighbours (KNN) (6) is an ML algorithm which can be employed for imputation, for example, it is implemented in 'caret' R package, 'knnImpute' function (7). The algorithm first finds k closest observations for which the parameter value has been observed, and then replaces the missing value with the neighbours' mean for

continuous values, or mode for categorical. Distance-based weighted averaging is also possible, in which lower weights are attributed to further located observations. As follows, the algorithm depends on the chosen distance measure and parameter k. Among the proposed distance measures the most popular is Euclidean, however others such as Minkowski metrics and correlation-based measures can be a better choice depending on the data (8,9). The method has been criticised for its low accuracy in high dimensional spaces, instability in smaller datasets, and slow computation times (10). In addition, imputing new observations would require keeping the entire training data for neighbours' selection, which is memory-intense and may raise privacy issues at implementation.

Random forests can be used for multivariate imputation (11). One of the most popular implementations is **MissForest** in 'missForest' R package (12). It follows an iterative procedure, in which missing values are first filled with initial guesses (e.g., mean), and RF is trained to predict missing values of the feature with the highest missingness based on the complete observations of this feature. From this RF model, predictions for imputed data are updated by averaging across observations in the final leaves in each tree, and across all trees in RF. Then, the missing values in the next feature are predicted, using the updated missing values of previously imputed features. The procedure is repeated until the difference in the matrices of the imputed values stops contracting (12). The main advantage is model's flexibility to handle complex and non-linear relationships, robustness to outliers, good performance in high missingness scenarios, and even some MNAR scenarios (13).. Further, it has been shown to outperform parametric regression models for clinical data (14). It can be argued that due to longer fitting times, other methods may be preferred for data with low missingness rate. Moreover, missForest cannot extrapolate, and produces missing values within the range of the training data, as predictions are done by averaging observed values in final leaves.

3 Multiple imputations (MI) using multivariate imputation by chained equations (MICE)

Single regression-based imputations have been criticised for their over-optimistic performance estimates, an inflation in alpha error and too-narrow confidence intervals

of the estimated parameters. Indeed, an imputation model works by finding relationships between the supplied features. It then replaces missing data with its plausible values in the context of this model, thus intensifying captured relationships. Therefore, using single imputation can create a false sense of certainty expressed in optimistic (too narrow) confidence intervals and undervalued standard deviations (2).

Addressing this critic, Rubin (15) proposed using multiple imputations (MI), in which a set of plausible versions of the complete data are generated to reflect the uncertainty of estimation. Van Buuren and Groothuis-Ousdshoorn (2000) proposed **multivariate imputation by chained equations method (MICE)** to produce MI. MICE is based on Gibbs sampling, which uses iterative process of updating plausible values for missing data until conversion. It is implemented in 'mice' R package (16) and extends the original MI method to handle several missing features simultaneously.

Conducting MI by MICE involves the following:

1. **Imputation:**
    - First, missing values are filled with initial guesses (e.g., mean values), or event random numbers.
    - Second, the vector of the first feature values across the subjects, $X_1 = (x_{11}, \ldots, x_{n1})^T$, is treated as an outcome and other features as predictors, and a regression model is fitted to the data with observed feature's value, $X_1^{obs} \sim X^0_{2,\ldots,} X^0_k$, where
      $X_1^{obs} = (x_{j1}$ for j such that x_{j1} is observed, or $R_{1j}=0)$, and
      $X^0_2 = (x_{j2}$, if x_{j2} observed, of its initial guess if not, for all j with $R_{1j}=0)$, …
      $X^0_k = (x_{jk}$, if x_{jk} observed, of its initial guess if not, for all j with $R_{1j}=0)$,
    - Third, a distribution of X_1 is estimated, given observed values X_1^{obs} and current guesses of other features $X^0_{2,\ldots,} X^0_k$., and X_1's missing values are updated by drawing the values randomly from this distribution.
    - Fourth, the procedure is repeated for each feature vector, X_j, j=1…p.
    - Steps 1-4 are looped into a cycle, each time starting with the updated values of the missing data from the previous step:

$$\hat{X}_1^t \sim \hat{P}\left(X_1 \middle| X_1^{obs}, \hat{X}_2^{t-1}, \ldots, \hat{X}_k^{t-1}\right)$$

…

$$\hat{X}_k^t \sim \hat{P}\left(X_k \middle| X_k^{obs}, \hat{X}_1^{t-1}, \ldots, \hat{X}_{k-1}^{t-1}\right) \qquad (2.78)$$

- The process stops when the imputing values converge, or a stopping criterion is met such as completing a given number of iterations or achieving a sufficient closeness of the imputed values between iterations.
- The imputation procedure is run *m* times such that different random values from the posterior distributions are drawn each time (16).

2. **Analysis**:
   - The study analyses are conducted for each imputed data separately.
3. **Pooling:**
   - Results of analysing all imputed data versions, such as performance measures and regression coefficients, are pooled using Rubin's rules (15). According to these rules, estimated means are averaged across the imputations to obtain pooled mean, while standard deviations are updated to accommodate between and within imputation variance (U and B below):

$$\text{Pooled mean } \hat{Q} = \frac{1}{m}\sum_{i=1}^{m} Q_i$$

$$\text{Pooled } Var(Q) = V1 + (1 + 1/m) \cdot V2,$$

$$V1 = \frac{1}{m}\sum_{i=1}^{m} V_i, \quad V2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(Q_i - \hat{Q}\right)^2 \qquad (2.79)$$

Here m – number of imputations, Q_i and V_i – estimates of statistic's mean and variance in the i^{th} imputed dataset (for instance, ROC-AUC of a model), V1- within imputation variance, V2 – between imputation variance. This paper extends Rubin's rules to other statistics such as likelihood ratio or correlation measures (17).

Optimal number of imputations.

As noticed by van Buuren (2018), even 2 imputations (*m*=2) would produce unbiased estimated and correct confidence intervals backed from the total variance in **equation 2.79**. However, being a simulation technique, MI estimates would be more precise with higher *m*. Rubin (1987) concluded that 2-10 imputations are enough for good point estimate, other studies mention numbers from 2 to 200 depending on the missingness

percentage and study aims (15,18–21). In particular, studies requiring high statistical power (e.g. aiming to detect small effect sizes), may require higher *m* (>20) for medium missingness (10-30%) (22). There are a number of simulation studies investigating which *m* allows 95% confidence interval of an estimated parameter to be within 10% of its theoretical value in 95% simulated cases. They showed that such *m* could be represented by a linear or quadratic function of the missingness rate: *m*=100 x percent missing, so m=10 for 10% missing data, *m*=30 for 30% (19), or *m* = 200 x (percent missing)^2, that is, *m*=2 for 10% missing and *m*=18 for 30% (21). Finally, practical choice of *m* can also depend on computation and memory resources (18).

Which features to impute and/or include to the imputation model.

Further question is which features should be included into the imputation model, and which features can or should be imputed at all. Most agree that imputing outcome or main predictor of interest should be avoided, and so is imputation of features with very high missingness rate (23,24). However, it may be argued that observations with missing outcome could still be used for prediction of other missing data at the imputation stage but excluded from the main analysis. If such imputed outcomes were included into the training data, prediction model would learn the relationships inferred by the imputation model, which would lead to biased results. On the other hand, using them for imputation preserves sample representativeness in which imputation model can be better informed on the missing predictor values.

Another nuanced question is whether to include (observed) outcome variable while imputing other features: its inclusion may improve imputation accuracy, but also reinforce relationships between the predictors and outcome and inflate prediction model performance. It has been shown, however, that excluding outcome from the imputation model is much more dangerous and could lead to substantial bias even under MCAR (25), as it had happened in the first version of the QDiabetes model, where unrealistic dependencies were observed in the resulting model (REF).

Different regression models can be used in MICE, that is, regression models in step 2 above. For instance, adaptation of MICE using random forests (RF-MICE) (26) includes building CART for predicting one feature from the others and updating imputed values by randomly choosing an observation from the final leave where this missing value falls.

RF-MICE may take longer to converge but can outperform linear regression MICE for non-linear data and even missForest (27).

Although MI by MICE is one of the preferred missing data handling method while developing a clinical prediction model (23,24), there are limitations. Apart from a trivial observation of an increased computation and coding complexity, it has been shown that chained equations return unstable imputations beyond expected statistical bounds in some scenarios (28). In addition, any particular choice of the underlying regression method for MI / MICE will have its corresponding limitations, such as linearity and normality assumptions for the linear regressions or long computation times for random forests (20).

4 Other methods for handling missing data

Imputation is not the only approach to dealing with incomplete data.

4.1 Maximum likelihood and expectation-maximisation (EM) algorithm

In maximum likelihood approach, instead of imputing missing data, the model estimates parameters that maximise the probability of seeing the data that have actually been observed. For that, missing values are viewed as random variables, usually from the joint normal distribution. Then, each subject contributes to the likelihood by the probability of seeing the its observed features marginalised (averaged) over the distribution of the unobserved features under MAR assumption (29). Formally, following Allison et al. (2010), for the data of *n* subjects represented by the *k* features with joint density function $f(x_1, \dots, x_k, \theta)$, and θ is parameters to estimate, ML for *complete* data is

$$L = \prod_{i=1}^{n} f_i(x_1, \dots, x_k, \theta) \qquad (2.80)$$

Then, if subject *i* misses the first two features, the likelihood of observing x_i is the probability of observing the rest (in MAR) is:

$$f_i^*(x_{i3}, \dots, x_{ik}, \theta) = \iint_{x1,x2} f_i(x_1, \dots, x_k, \theta) dx_1 dx_2 \qquad (2.81)$$

Then, ML for incomplete data would look like the following, assuming *m* subjects had no missing data and *(n-m)* were incomplete:

$$L = \prod_{i=1}^{m} f_i(x_{i1}, \dots, x_{ik}, \theta) \prod_{j=m+1}^{n} f_j^*(X_{j\,observed}, \theta) \quad (2.82)$$

Equation 2.82 can be solved using EM algorithm (30), where so called expectation (E) and maximisation (M) steps are alternated. First, initial guess for θ is set, and missing values that maximise **equation 2.82** are computed (E step), then, given these values, parameters θ is updated by maximising likelihood by θ (M step). The procedure is iterated until the parameters stop changing.

Maximum likelihood was considered to be the main alternative to MI (29,31): it is asymptotically more efficient than MI, reproducible (gives the same estimates every time, unlike MI), and has less hyperparameters than MI (31). While this is true, the result of this method is a regression with estimated parameters and hence this regression can predict new cases with complete data, but not those with missing values, which limits its applicability to prediction modelling. Moreover, it is computationally expensive and often not converge in high missingness scenarios, which perhaps explains its low popularity in published studies. For example, no studies in a recent UK clinical prediction models review used this method to handle missing data (32).

4.2 Models with embedded algorithms for missing data

Somewhat similar to the maximum likelihood approach, a number of machine learning methods such as CART, RF, XGBoost, and neural networks, have been extended to handle incomplete data directly. So, instead of separating imputation model and regression algorithm used for the main analyses, regression models can be empowered with an embedded missing data algorithm be fitted to the incomplete data.

Although different in details, all tree-based methods (CART, RF, XGBoost) view missing values as a special category while perform data splits, and compare reduction in the loss function for attributing missing data to the left or right node (11). The resulting model would learn the patterns in both observed and unobserved data, and hence able to predict for unseen samples with missing data as well as for the training sample.

Neural networks (NN) are algorithms inspired by brain neural networks which involve many interconnected neurons and many of these algorithms can handle missing data intrinsically. Some fill incomplete inputs by selecting the values with minimise learning

criterion, others use so called context encoders but require complete training data to handle missing values thereafter (33). That said, NN are highly customised methods; granted that most deep learning missing data approaches have been developed for imaging or big multidimensional data, its generalisability to clinical settings is difficult to assess.

4.3 Emerging missing data methods for prediction modelling

Some argue that existing missing data handling methods should be further adapted to prediction modelling (34). Indeed, most of existing methods were developed in the context of explanatory research, where accurate and unbiased estimates of model parameters are needed. However, this is not directly relevant for prediction tasks, in which prediction accuracy is of the main concern, and biased or embedded missing data algorithms can be a valid alternative. Further, missingness patterns themselves can have predictive power. For example, knowing that a medical test was not conducted for a certain patient may represent latent information which prompted clinicians' decision to not order the test. In this case, the fact of missingness may be predictive of the outcome irrespective of which values the test could have had and can be introduced to the model by adding missingness indicators. To note, embedded handling of incomplete data by ML algorithms described above may also be considered as models learning both from the observable values and missingness patterns. The main concern of this approach is to let the model learn non-generalisable missingness patterns to other settings, which should be assessed on a case by case basis (34).

Agreeing with Sperrin et al, it seems that a non-negligible portion of missing data in clinical samples is NMAR and simplifying it to non-missing such as assuming no disease if there is no diagnosis, or MAR may not always be optimal for clinical prediction models. Instead, underlying unobservable mechanisms may be guessed or modelled with a missing data indicator, which has shown to be effective in some settings (35–37).

5 Conclusion remarks on missing data

Most of clinical data is incomplete, especially in observational and/or longitudinal studies. Situations, when participants miss follow-ups or drop out early from study, are

frequent, while book-keeping and digitisation of clinical records is never flawless. On the other hand, incorrect and biased predictions can have serious consequences for patient care and clinical decision-making. Therefore, handling missing data in clinical research and prediction modelling is an essential part of the analysis. The importance of proper justification and reporting of the missing data handling methods has been highlighted in many modern statistical guidelines such as Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) (38), (39), or (40). In practice, however, adherence to these recommendations has been rather poor as recent review of the UK clinical prediction models suggests (32). Out of 23 studies developing the models and 210 external validation papers, around 50% did not report the way missing data was handled, 40% used complete cases analysis, 5% used various single imputation models, and only 5% used MI (32).

That said, the decision on which approach to take is never straightforward, and a combined strategy is often recommended (23,29,39). Some observations or features can be excluded if percent missing is high, while the rest is imputed by multiple or single imputation driven by the sample size, missingness patterns, computational resources, and software availability.

References

References

1. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–592.
2. Buuren S. Flexible imputation of missing data. CRC press; 2018.
3. Kenward MG, Carpenter J. Multiple imputation: current perspectives. Stat Methods Med Res. 2007 Jun;16(3):199–218.
4. Austin PC, White IR, Lee DS, van Buuren S. Missing data in clinical research: A tutorial on multiple imputation. Can J Cardiol. 2021 Sep;37(9):1322–1331.
5. Little RJA, Rubin DB. Statistical Analysis with Missing Data. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2002.
6. Altman NS. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. Am Stat. 1992 Aug;46(3):175–185.
7. Kuhn M. Building Predictive Models in *R* Using the caret Package. J Stat Softw. 2008 Oct 11;28(5).
8. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001 Jun;17(6):520–525.
9. Karayiannis NB, Randolph-Gips MM. Non-Euclidean c-means clustering algorithms. IDA. 2003 Nov 17;7(5):405–425.
10. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. Springer; 2009.
11. Tang F, Ishwaran H. Random forest missing data algorithms. Stat Anal Data Min. 2017 Dec;10(6):363–377.
12. Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012 Jan 1;28(1):112–118.
13. Cutler A, Cutler DR, Stevens JR. Tree-Based Methods. In: Li X, Xu R, editors. High-Dimensional Data Analysis in Cancer Research. New York, NY: Springer New York; 2009. p. 1–19.
14. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. Am J Epidemiol. 2014 Mar 15;179(6):764–774.
15. Rubin DB, editor. Multiple imputation for nonresponse in surveys. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 1987.
16. van Buuren S, Groothuis-Oudshoorn K. mice : Multivariate Imputation by Chained Equations in *R*. J Stat Softw. 2011;45(3).

17. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol. 2009 Jul 28;9:57.

18. van Buuren S. Flexible imputation of missing data, second edition. Second edition. | Boca Raton, Florida : CRC Press, [2019] |: Chapman and Hall/CRC; 2018.

19. Bodner TE. What Improves with Increased Missing Data Imputations? Structural Equation Modeling: A Multidisciplinary Journal. 2008 Oct 22;15(4):651–675.

20. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011 Feb 20;30(4):377–399.

21. von Hippel PT. How Many Imputations Do You Need? A Two-stage Calculation Using a Quadratic Rule. Sociol Methods Res. 2018 Jan 18;004912411774730.

22. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev Sci. 2007 Sep;8(3):206–213.

23. Steyerberg EW. Clinical prediction models: A practical approach to development, validation, and updating. Cham: Springer International Publishing; 2019.

24. Petersen I, Welch CA, Nazareth I, Walters K, Marston L, Morris RW, et al. Health indicator recording in UK primary care electronic health records: key implications for handling missing data. Clin Epidemiol. 2019 Feb 11;11:157–167.

25. Moons KGM, Donders RART, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006 Oct;59(10):1092–1101.

26. Doove LL, Van Buuren S, Dusseldorp E. Recursive partitioning for missing data imputation in the presence of interaction effects. Comput Stat Data Anal. 2014 Apr;72:92–104.

27. Shah A, Bartlett J, Carpenter J, Nicholas O, Hemingway H. Comparison of parametric and Random Forest MICE in imputation of missing data in survival analysis. 2022 Dec 4 [cited 2023 Mar 27]; Available from: https://cran.r-project.org/web/packages/CALIBERrfimpute/vignettes/simstudy_survival.pdf

28. Chen S-H, Ip EH. Behavior of the gibbs sampler when conditional distributions are potentially incompatible. J Stat Comput Simul. 2015;85(16):3266–3275.

29. Johnson DR, Young R. Toward Best Practices in Analyzing Datasets with Missing Data: Comparisons and Recommendations. J Marriage and Family. 2011 Oct;73(5):926–945.

30. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the *EM* Algorithm. Journal of the Royal Statistical Society: Series B (Methodological). 1977 Sep;39(1):1–22.

31. Allison PD. Survival analysis using SAS: a practical guide. books.google.com; 2010.

32. Tsvetanova A, Sperrin M, Peek N, Buchan I, Hyland S, Martin GP. Missing data was handled inconsistently in UK prediction models: a review of method used. J Clin Epidemiol. 2021 Sep 11;140:149–158.

33. Śmieja M, Struski Ł, Tabor J. Processing of missing data by neural networks. Advances in neural 2018;

34. Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. J Clin Epidemiol. 2020 Sep;125:183–187.

35. Sharafoddini A, Dubin JA, Maslove DM, Lee J. A new insight into missing data in intensive care unit patient profiles: observational study. JMIR Med Inform. 2019 Jan 8;7(1):e11605.

36. Seaman S, White I. Inverse Probability Weighting with Missing Predictors of Treatment Assignment or Missingness. Communications in Statistics - Theory and Methods. 2014 Aug 18;43(16):3499–3515.

37. Fletcher Mercaldo S, Blume JD. Missing data and prediction: the pattern submodel. Biostatistics. 2020 Apr 1;21(2):236–252.

38. Collins GS, Reitsma JB, Altman DG, Moons KGM, TRIPOD Group. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD Group. Circulation. 2015 Jan 13;131(2):211–219.

39. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014 Aug 1;35(29):1925–1931.

40. Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies. Int J Med Inform. 2021 Sep;153:104510.