Journal of Clinical Epidemiology 125 (2020) 183−187

Journal of Clinical Epidemiology

COMMENTARY

Missing data should be handled differently for prediction than for description or causal explanation

Matthew Sperrin*, Glen P. Martin, Rose Sisk, Niels Peek

Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK

Accepted 18 March 2020; Published online 12 June 2020

Abstract

Missing data are much studied in epidemiology and statistics. Theoretical development and application of methods for handling missing data have mostly been conducted in the context of prospective research data and with a goal of description or causal explanation. However, it is now common to build predictive models using routinely collected data, where missing patterns may convey important information, and one might take a pragmatic approach to optimizing prediction. Therefore, different methods to handle missing data may be preferred. Furthermore, an underappreciated issue in prediction modeling is that the missing data method used in model development may not match the method used when a model is deployed. This may lead to overoptimistic assessments of model performance. For prediction, particularly with routinely collected data, methods for handling missing data that incorporate information within the missingness pattern should be explored and further developed. Where missing data methods differ between model development and model deployment, the implications of this must be explicitly evaluated. The trade-off between building a prediction model that is causally principled, and building a prediction model that maximizes the use of all available information, should be carefully considered and will depend on the intended use of the model.

Keywords: Clinical prediction models; Missing data; Multiple imputation; Prognostic model; Routinely collected data; Model performance

1. Background

The study of missing data is well established in epidemiology and statistics. However, the focus of almost all missing data literature to date has been on its implications for parameter estimation, which is relevant in studies whose objectives are description or causal explanation [1]. For this, one wishes to handle missing data in such a way as to minimize or eliminate bias in particular parameters of interest and also maximize precision in their estimation. In prediction studies, however, we wish to maximize predictive accuracy, where bias and precision in specific parameter estimates is not a *direct* concern, and instead one emphasizes pragmatism in building models that optimize for prediction. Therefore, optimal strategies for handling missing data may differ.

Classically, missing data are divided into different generative scenarios or assumptions [2]. Missing completely at random (MCAR) occurs when the probability of a particular observation being missing is independent of all other data. Missing at random (MAR) is when the probability of missing depends only on observed data, whereas missing not at random (MNAR) covers the remaining scenario that the probability of missing depends on both observed and missing data. In general, MCAR and MAR are less problematic, whereas MNAR can often cause bias in effect estimates that cannot be overcome. This MCAR/MAR/MNAR classification is, however, motivated by the need to estimate parameters of interest and is less suitable for prediction, where the important consideration is whether the missingness is related to the outcome [3]. In turn, much of the focus with regard to identifying appropriate methods to handle missing data is on ensuring that bias is not introduced into parameter estimates. For example, multiple imputation is primarily motivated from the fact that it gives unbiased estimators under MAR, provided the imputation model is correctly specified [4]. In the

Conflicts of interest and source of funding: This work was supported by the Alan Turing Institute partnership project ‘Predictive Healthcare’, by the National Institute for Health Research Manchester Biomedical Research Center, and by the Medical Research Council (Grant number MR/*N*013751/11).

* Corresponding author. University of Manchester, Vaughan House, Manchester M13 9PL, UK. Tel.: +44 (0) 161 3067629.

E-mail address: matthew.sperrin@manchester.ac.uk (M. Sperrin).

https://doi.org/10.1016/j.jclinepi.2020.03.028

What is new?

Key findings

- To date, missing data research has primarily been conducted in the context of parameter estimation, which is not directly relevant for prediction.
- For prediction, a more pragmatic approach to missing data handling may be warranted, to optimize predictive accuracy.
- Multiple imputation may not be the best approach to handle missing data when building prediction models, especially when missingness can be informative.

What this adds to what was known?

- The optimal approach to handle missing data for prediction may differ from the optimal approach to handle missing data for other objectives such as causal explanation.
- The missing data approach used in prediction model development may not match the approach used when a model is deployed.

What is the implication and what should change now?

- When developing prediction models, one should consider whether missing data patterns are informative and consider using missing pattern approaches if so.
- However, the missingness patterns, and the mechanisms underlying them, are likely to change over time, so close monitoring is required, and more research is required in this area.
- Validation of prediction models should replicate the missing data approach that would be used at 'prediction time' where this differs from the approach used at 'development time'.

predictive setting, however, causal effects are often less of a concern.

Moreover, research is increasingly conducted using routinely collected data such as electronic health records, in which prevalence of missing data is likely to be high, and the mechanism is likely MNAR [5]. This is a challenge particularly for causal inference (that we do not consider here), but perhaps an opportunity for predictive modeling because the missing pattern itself could convey information that improves prediction. In this article, we discuss the challenges and opportunities of missing data in the context of prediction, particularly with real-world data. We emphasize that this is a distinct problem, with different solutions, compared with missing data for causal inference.

2. Exploiting missing data patterns in prediction

When using 'traditional' research data, such as a prospective cohort study, missingness often has relatively low prevalence (assuming the study is well designed and conducted) and assuming that data are MAR is generally seen to be reasonable. Hence, missing data have been viewed as little more than a nuisance, and the recommended method in prediction modeling has been to use multiple imputation, primarily to maximize efficiency and reduce biases [6].

Increasingly, we build prediction models using routinely collected 'real-world' data. For example, QRISK is developed using electronic health records [7]. Here, missingness or presence of a particular predictor may be highly informative of outcome, and thus incorporation of this information in a model may improve prediction. For example, if a particular test has been carried out on a patient, then the perceived need for the test tells us something about the overall condition of the patient. In particular, it may provide information about factors related to the outcome that are unmeasured, that motivated a clinician's decision to request a test. Hence, the presence of the test, independent of the test result, can act as a proxy for this latent information [8].

There are many different situations where missingness in predictors might be informative in this way, besides presence or absence of a particular test. First, we may have less information on patients who have fewer contacts with a clinician. This may be because the patient is healthy, but could be that they are too unwell to attend an appointment. Second, symptoms may be unrecorded in a patient's record because they are deemed irrelevant. Third, missingness patterns in rich longitudinal data may be informative. For example, if a patient's heart rate is being tracked, they may not wear the heart rate monitor when they are sleeping, whereas for another patient, the heart rate may not be measured accurately because the patient is exercising vigorously. These reflect MNAR patterns and hence the missingness in the predictors itself has the potential to be predictive of the outcome (although note that predictors can be MNAR but not necessarily predictive of the outcome).

As an example of early exploration of the incorporation of missing patterns into prediction, Van der Heijden et al. [9] considered different imputation techniques in the context of building a multivariable prediction model. They consider both multiple imputation and a simple version of the missing indicator method with mean imputation [9]. They find that the missing indicator method resulted in a predictive model with the highest area under curve, showing the value in incorporating missing indicators for

prediction. Interestingly, the authors call this an 'overestimated receiver operating characteristic area' and maintain that the missing indicators have 'no diagnostic importance', highlighting an initial reluctance in the prediction modeling community to embrace the potential value of missingness.

More recently, Sharafoddini et al. [10] built a model to predict short-term mortality in the intensive care unit setting. They used multiple imputation, but also considered including missing indicators in the prediction model. Primarily these missing indicators captured whether particular tests had been ordered. Sharafoddini et al. [10] found that inclusion of the missing indicators led to substantial improvements in prediction. Indeed, using the missing indicators alone had good predictive performance. The combination of multiple imputation and missing indicators is emerging as a useful technique in a range of settings [8,11−13]. Although missing indicators are known to introduce bias in estimation of causal effects even under MCAR [14], this does not preclude their inclusion in prediction models.

The very notion of missing data is not always a helpful one, particularly in routinely collected data, such as in electronic health records, where there is no strict protocol governing when data are observed. In these cases, it can be informative to explicitly model the times at which data are observed: such a model is often termed the 'visit process' or 'observation process' [15]. For example, a patient's blood pressure (or any continuously evolving biomarker or risk factor) can be measured, in principle, at any time. The times at which it is measured—the observation process—are likely to convey important information, over and above the actual result of the measurement; this is equivalent to MNAR. Methods are emerging to handle such observation processes, although as with missing data, most of these focus on producing unbiased parameter estimates [15]. Although some methods are emerging that allow the observation process to be exploited for prediction [16], there is little evidence to date that informative observation is considered in prediction models in practice [17].

3. Missing data at model development vs. model validation vs. model deployment

An underappreciated issue in predictive research is the distinction between handling missing data at model development time, at model validation time, and at model deployment time (or prediction time) [18]. For example, multiple imputation (with the outcome included in the imputation model, as recommended [19]) is often used during model development and validation [20]. However, performing multiple imputation at model deployment time is difficult or infeasible for a number of reasons. First, one cannot use the outcome to inform the imputation because the outcome is clearly unknown at the prediction time. Second, substantial effort and data are required to calculate the imputed values required for missing data at this prediction time. This includes the original development data, to which the new data must be appended, and imputation rerun [13,21]. Indeed, when applying the model in practice, simpler one-step imputation methods might be used. Therefore, external validation studies that resolve missing data using multiple imputation may overestimate the performance of deployed prediction models [13]. We therefore recommend that validation should include an evaluation of performance when missing data are handled using the same methods that will be used when the model is deployed.

A promising alternative to multiple imputation that overcomes some of these challenges, considered by Saar-Tsechansky and Provost [22], and recently revisited by Fletcher-Mercaldo et al. [13], is the missing pattern method. Here, a separate model is fit for each missingness configuration, based on the data in the development set that match that missingness configuration. This means that, at prediction time, the prediction can be provided quickly without the need to rerun imputations. The downside is the high complexity and potential overfitting at model development time, for p predictors 2^p models are required, and clearly many of these may be based on few or no observations. However, Fletcher-Mercaldo et al. [13] proposed ways to overcome these limitations, such as parallel computing at development time and borrowing strength across different missingness patterns that are similar but represented by few or no observations. They found that their missing pattern method was superior to multiple imputation in terms of squared error loss in predicting a continuous outcome. A related method is to fit a model for each missingness pattern using all of the development data, rather than restricting to the subset of the development data with the matching missing pattern [21]. However, this makes stronger assumptions about the missingness mechanisms and has poor performance [13,21].

A further problem with the distinction between handling missing data at development and prediction time is that the missingness mechanism is highly likely to change once a predictive model is being used in practice. Trivially, if a clinician wishes to use a predictive model, they are far more likely to measure and collect the variables that are required as inputs. Therefore, missingness is likely to be lower than in the development data. When missingness does still exist, however, it is likely to reflect a different mechanism because the underlying observational process has changed [23].

There are strong parallels between this discussion and recent investigations in measurement error, particularly that when predictors are measured differently between the development, validation, and deployment of a prediction model (measurement heterogeneity), this will affect the performance of the model [24−26].

4. Prediction models may benefit from causal consideration

Although the use of missing indicators and missing patterns may seem appealing, there is a natural concern over generalizability and external validity of a prediction model that relies heavily on such features. This is partly because, as discussed above, the information conveyed by a missing value will differ between development and prediction time. However, it is also highly likely that, independent of the presence or absence of a predictive model, the missingness mechanisms are highly context dependent, so such a model may have poor geographical or temporal generalizability [27]. It is also partly because, although causal relationships are not of interest per se in predictive modeling, it is clear that causal relationships will generalize better across settings. Missingness patterns are unlikely to have a direct causal effect on outcome. Therefore, a predictive model that introduces bias into causal relationships, or relies heavily on missingness patterns, may be less generalizable. Moreover, incorporation of causal information in a predictive model allows for counterfactual prediction [28]. This has been shown to allow calculation of treatment naive risk in presence of treatment drop-in [29] and potentially even to compare different treatment strategies [30]. Therefore it is particularly useful whenever a decision is to be made on the basis of a prediction model [28].

Many of the above issues can, however, be overcome. Methods to handle missingness that both minimize bias and exploit missing pattern information are available [13]. Moreover, with the increased digitization of health care systems, the opportunity exists to create dynamic predictive models that rapidly correct for any error introduced by inclusion of noncausal effects or change in missingness mechanisms across time and space [31]. The emphasis on incorporating causal information in a prediction model will also depend on the purpose of the model. Where generalizability across settings is required, or counterfactual prediction calculations are warranted, it should be prioritized. In a model tailored to a particular setting, or used for risk adjustment or audit, it should be de-emphasized.

Creating models that trade off the need for accurate predictions, with the need to produce accurate counterfactual contrasts remains an underexplored area of research.

5. Conclusion and recommendations

It is important to acknowledge the distinction in the implications of missingness in models designed to predict and models designed to address other objectives such as description or causal explanation. The way that missingness should be handled, while sharing common features, should be different, with the similarities dependent on the intended use of the prediction model. For prediction, bias in effect estimates is less important, but should not be forgotten about entirely, particularly when aiming to use the model to aid decision making. Missingness and observation patterns can convey useful information and thus improve predictive accuracy, particularly in routinely collected data, but should be incorporated and interpreted with caution. Furthermore, the distinction between how missingness can and should be handled in model development, and in model deployment, is crucial. If missing data are to be handled differently in development and deployment, the implications need to be understood.

To conclude, missing pattern information in covariates can and should be used to increase the accuracy of prediction models. In particular, missing indicators are reasonable predictors to include (with caution) as a simple way of incorporating the information that missingness may provide. As with all predictive models, those that incorporate missing information should be closely monitored, and revised, preferably in a dynamic scheme, as missingness mechanisms and other pertinent factors change over time and space. Further research is required concerning how missing data or observation processes can be exploited in prediction, but without adversely impacting on generalizability and robustness of the predictive model.

CRediT authorship contribution statement

Matthew Sperrin: Conceptualization, Writing - original draft, Writing - review & editing. **Glen P. Martin:** Conceptualization, Writing - review & editing. **Rose Sisk:** Conceptualization, Writing - review & editing. **Niels Peek:** Conceptualization, Writing - review & editing.

References

[1] Shmueli G. To explain or to predict? Stat Sci 2010;25:289−310.
[2] Rubin DB. Inference and missing data. Biometrika 1976;63:581−92.
[3] Ding Y, Simonoff JS. An investigation of missing data methods for classification trees applied to binary response data. J Mach Learn Res 2010;11:131−70.
[4] Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006;59:1087−91.
[5] Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit Transl Bioinform 2010;2010:1−5.
[6] Steyerberg EW, van Veen M. Imputation is beneficial for handling missing data in predictive models. J Clin Epidemiol 2007;60:979.
[7] Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ 2017;357:j2099.
[8] Choi J, Dekkers OM, le Cessie S. A comparison of different methods to handle missing data in the context of propensity score analysis. Eur J Epidemiol 2019;34:23−36.
[9] van der Heijden GJMG, Donders T, Stijnen T, Moons KGM. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 2006;59:1102−9.

[10] 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;7:e11605.
[11] Qu Y, Lipkovich I. Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. Stat Med 2009;28:1402−14.
[12] Seaman S, White I. Inverse probability weighting with missing predictors of treatment assignment or missingness. Commun Stat Methods 2014;43:3499−515.
[13] Fletcher Mercaldo S, Blume JD. Missing data and prediction: the pattern submodel. Biostatistics 2018;21:236−52.
[14] Groenwold RHH, White IR, Donders ART, Carpenter JR, Altman DG, Moons KGM. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. CMAJ 2012;184:1265−9.
[15] Pullenayegum EM, Lim LS. Longitudinal data subject to irregular observation: a review of methods with a focus on visit processes, assumptions, and study design. Stat Methods Med Res 2014. 0962280214536537.
[16] Alaa AM, Hu S, van der Schaar M. Learning from clinical judgments: semi-markov-modulated marked Hawkes processes for risk prognosis. In: Proc 34th Int Conf Mach Learn.
[17] Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2016;24:198−208.
[18] Wood AM, Royston P, White IR. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. BMJ 2015;57:614−32.
[19] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393.
[20] Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. BMJ 2010;340:c2442.
[21] Janssen KJM, Vergouwe Y, Donders ART, Harrell FE, Chen Q, Grobbee DE, et al. Dealing with missing predictor values when applying clinical prediction models. Clin Chem 2009;55:994−1001.
[22] Saar-Tsechansky M, Provost F. Handling missing values when applying classification models. J Mach Learn Res 2007;8:1623−57.
[23] Peek N, Sperrin M, Mamas M, van Staa T-P, Buchan I. Hari seldon, QRISK3, and the prediction paradox. BMJ 2017;357:j2099.
[24] Luijken K, Groenwold RHH, Van Calster B, Steyerberg EW, van Smeden M. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: a measurement error perspective. Stat Med 2019;38:3444−59.
[25] Pajouheshnia R, van Smeden M, Peelen LM, Groenwold RHH. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. J Clin Epidemiol 2019;105:136−41.
[26] Luijken K, Wynants L, Smeden M van, Calster B Van, Steyerberg EW, Groenwold RHH, et al. Changing predictor measurement procedures affected the performance of prediction models in clinical examples. J Clin Epidemiol 2020;119:7−18.
[27] Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013;10:e1001381.
[28] Hernán MA, Hsu J, Healy B. Data science is science's second chance to get causal inference right: a classification of data science tasks. Chance 2018;32:42−9.
[29] Sperrin M, Martin GP, Pate A, Van Staa T, Peek N, Buchan I. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. Stat Med 2018;37:4142−54.
[30] Schulam P, Saria S. Reliable decision support using counterfactual models. In: Advances in Neural Information Processing Systems. ArXiv E-Prints; 2017:1697−708.
[31] Jenkins DA, Sperrin M, Martin GP, Peek N. Dynamic models to predict health outcomes: current status and methodological challenges. Diagn Progn Res 2018;2:23.