WILEY

CRITICAL COMMENTARY

Is the Most Likely Value Also the Best Imputation?

Stef van Buuren[1,2] | Hanne I. Oberman[1]

[1]Department of Methodology and Statistics, Utrecht University, Utrecht, the Netherlands | [2]Department of Child Health, Netherlands Institute for Applied Scientific Research TNO, Leiden, the Netherlands

Correspondence: Stef van Buuren (s.vanbuuren@uu.nl)

1 | Introduction

Missing data are an unavoidable feature of registry-based pregnancy studies. In their article, Sakai and colleagues [1] study approaches for recovering incomplete maternal data in the Norwegian Medical Birth Registry. They compared "longitudinal methods", which borrow information across repeated pregnancies from the same mother by tailored data operations (see their table S1), with an approach they label "multiple imputation", which does not utilize the same-mother pairs explicitly. They evaluated these methods by masking observed values and comparing reconstructed entries with the true ones, reporting the root mean squared error (RMSE) for continuous variables and agreement measures for categorical ones. Based on these metrics, they concluded that longitudinal methods generally outperformed multiple imputation.

The study deserves credit for highlighting the value of longitudinal information in repairing registry data. While such methods are only available for women with multiple pregnancies, they can be highly effective. Deterministic updates should be preferred over multiple imputation when the value can be deduced with certainty. At the same time, the analysis inadvertently reinforces some common misconceptions about multiple imputation. In what follows, we discuss three major pitfalls in their approach.

2 | Problem 1: Not Really Multiple Imputation

A first concern is that what Sakai et al. describe as multiple imputation is, in fact, *single imputation in disguise*. The authors generated 50 imputed datasets, but instead of analysing them separately and combining results according to Rubin's rules, they averaged the imputed values across datasets. For categorical variables, they even rounded these averages to integers. The result is a singly imputed dataset, stripped of the variation that defines multiple imputation.

This approach to multiple imputation is contrary to the philosophy that underlies the method. Multiple imputation is not about producing the most likely replacement for each missing cell, but about reflecting the uncertainty surrounding those missing values [2]. The strength of multiple imputation lies precisely in the variability across imputations, which allows standard errors and confidence intervals to incorporate the missing-data uncertainty. By collapsing the imputations into a single best guess, the method used by Sakai et al. cannot benefit from this mechanism.

This distinction is more than a technicality. A singly imputed dataset typically underestimates standard errors and overstates the precision of results [3]. While the dataset may appear plausible at face value, inferences drawn from it can be misleading. Methodological guidance is clear on this point: collapsing multiple imputations into one is not recommended practice [4]. It is therefore important to recognize that the study does not actually evaluate multiple imputation as it is properly understood and applied.

3 | Problem 2: Ignoring Matched Pairs

A second limitation is the way Sakai et al. implemented their imputation strategy. The authors present longitudinal and multiple imputation methods as two distinct and mutually exclusive alternatives: one exploits repeated pregnancies from the same

mother; the other does not. In practice, however, these approaches can be combined.

Multiple imputation is highly flexible and can incorporate donor information from matched records. Figure 1 illustrates a simple strategy: merging the index pregnancy (where the value is missing) with the reference pregnancy (where the value is observed) into a single record. This structure allows models to capture associations not only within the index record but also across the matched pair. When the reference pregnancy provides a value that is logically determined, such as parity or maternal age differences, this can be computed directly through passive imputation, for example, using the *mice* package in R [5].

The advantage is that we retain the richness of the same-mother linkage while also capturing the uncertainty for variables where prediction is less certain. In other words, longitudinal information and multiple imputation are not competing paradigms but complementary tools. By not exploiting this possibility, Sakai et al. set up a false contrast that underplays the capacity of multiple imputation to integrate longitudinal information.

4 | Problem 3: Reconstruction Accuracy as Evaluation

The third, and perhaps most fundamental, issue lies in the choice of evaluation metrics. Sakai et al. assessed performance by comparing imputed values directly to observed ones, reporting RMSE for continuous variables and agreement measures for categorical ones. At first sight, this seems intuitive: if the reconstructed value is close to the original, the method must be good.

The problem is that minimising RMSE and maximising agreement systematically favours imputations that reproduce the most likely value in every missing cell. This may be appropriate for applications where each record is used individually, such as estimating a patient's weight in clinical care. But in epidemiological research, imputing the most likely value is problematic. It underestimates the variance, amplifies associations among variables, and distorts regression weights. The downstream effect is

overly precise and potentially biased estimates, even if the cell-level RMSE looks small [4].

This property is not a minor quirk; it reflects a fundamental distinction between prediction and inference. Multiple imputation was developed precisely to avoid this trap by propagating uncertainty through variation across imputations and combining results with Rubin's rules. Evaluation should therefore focus on *analysis-level validity* rather than *data-level accuracy*. A method can perform poorly on RMSE *yet yield unbiased regression estimates, or conversely, excel on RMSE* while producing misleading inferences.

By using reconstruction accuracy as their primary criterion, Sakai et al. inadvertently reinforced a widespread misunderstanding about what constitutes a "good" imputation. The risk is that readers may conclude that single imputation is preferable when, in fact, it undermines the very goals of epidemiological analysis.

5 | Implications for Perinatal Epidemiology

These methodological issues have real consequences for research in perinatal epidemiology. Registry-based studies rarely focus on predicting individual values; their purpose is to estimate associations across large populations. When expected values are substituted for missing values, the resulting datasets may appear complete and plausible, but they conceal the uncertainty that accompanies missingness. Analyses conducted on such datasets will tend to exaggerate associations, underestimate standard errors, and produce overconfident conclusions.

For practitioners, the practical lesson is twofold. First, multiple imputation should not be dismissed based on results from an incorrectly implemented version. Properly applied multiple imputation can easily accommodate longitudinal information. Second, the choice of evaluation metric must match the intended use of the data. Reconstruction accuracy may have value in clinical prediction settings, but for epidemiological inference, the relevant criteria are bias, coverage, and efficiency of the scientific quantities of interest [6, 7].
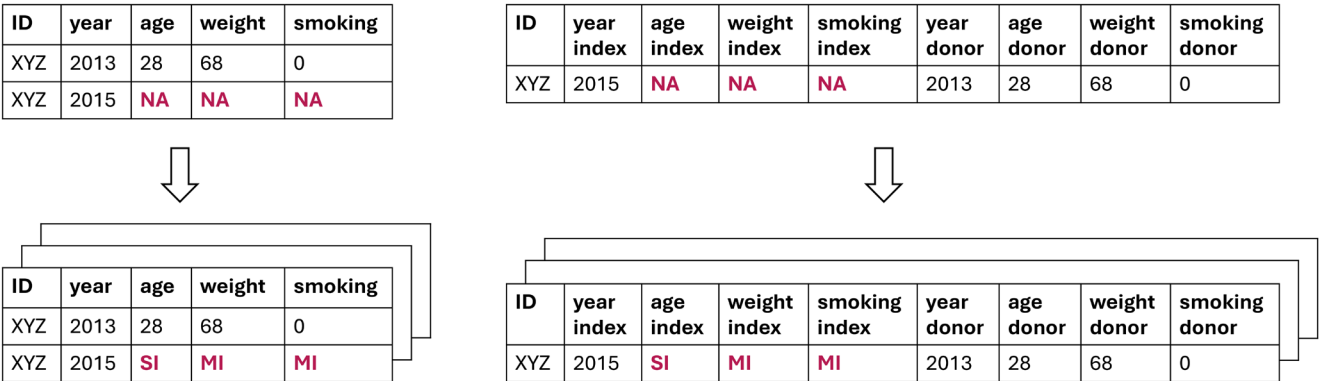
FIGURE 1 | Long (left) and wide (right) formats of longitudinal data. In the long format, missing values can be addressed with hierarchical models that account for the clustering of pregnancies within mothers. A more practical alternative is the wide format, where information from repeated pregnancies is placed on a single record. Deterministic single imputation (SI) can be applied to fields that can be calculated with certainty (e.g., maternal age), while multiple imputation (MI) is reserved for variables where uncertainty remains.

In short, abandoning multiple imputation in favor of deterministic copying methods would be premature. The better path is to integrate longitudinal insights into a multiple imputation framework that preserves uncertainty while leveraging the richness of registry data.

6 | Remaining Gaps and Future Directions

The study by Sakai et al. highlights a more general challenge: misconceptions about multiple imputation remain widespread. Many researchers equate "good" imputation with reproducing the missing values as closely as possible, overlooking that the ultimate goal is valid inference. Training materials, tutorials, and methodological guidance should make this distinction clearer.

Future research should demonstrate how registry-based strategies, such as copying from a prior or subsequent pregnancy, can be embedded in properly executed multiple imputation procedures. Such approaches would allow deterministic updates for variables known with certainty, while preserving variability for those where uncertainty remains. Simulation studies can illustrate how these hybrid methods perform under different missing-data mechanisms, evaluating bias, coverage, and efficiency rather than relying solely on reconstruction accuracy.

More broadly, perinatal epidemiology would benefit from collaborations between registry experts and methodologists to design imputation strategies that are both practical and statistically principled. Only then can we fully exploit the unique strengths of Nordic and other population registries without compromising analytical validity.

7 | Conclusions

Sakai et al. provide a valuable service by drawing attention to the problem of missing maternal data in registry-based studies and by showcasing the potential of repeated pregnancy records. Yet their work also illustrates three persistent misconceptions: what they implemented was not multiple but single imputation; they missed the opportunity to combine matched pregnancy records with multiple imputation; and they evaluated methods using reconstruction accuracy, a flawed criterion for epidemiological research.

Taken together, these issues caution against abandoning multiple imputation based on their findings. The more promising path is to integrate longitudinal information within a proper multiple imputation framework, thereby preserving uncertainty while exploiting the richness of registry data.

References

1. T. Sakai, H. Nordeng, and M. M. H. J. van Gelder, "Longitudinal Methods Versus Multiple Imputation to Infer Missing Maternal Data in Registry-Based Pregnancy Studies," *Paediatric and Perinatal Epidemiology* 0 (2025): 1–14.

2. D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys* (Wiley, 1987).

3. R. J. A. Little and D. B. Rubin, *Statistical Analysis With Missing Data*, 3rd ed. (Wiley, 2020).

4. S. van Buuren, *Flexible Imputation of Missing Data*, 2nd ed. (Chapman & Hall/CRC, 2018).

5. S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software* 45, no. 3 (2011): 1–67.

6. T. P. Morris, I. R. White, and M. J. Crowther, "Using Simulation Studies to Evaluate Statistical Methods," *Statistics in Medicine* 38 (2019): 2074–2102.

7. H. I. Oberman and G. Vink, "Toward a Standardized Evaluation of Imputation Methodology," *Biometrical Journal* 66 (2024): 2200107.