# On the Use of Multiple Imputation Approach in Pediatric Research

Sixia Chen, PhD, and Michael P. Anderson, PhD

In recent study from Salmon et al reported in *The Journal*, the investigators assessed the association between clinical chorioamnionitis and neurodevelopmental disorders at 5 years of age in children born preterm.[1] The authors did not find a significant association between the 2 variables based on epidemiological study on small gestational ages (EPIPAGE 2), a national, population-based cohort study of children born before 35 weeks of gestation in France in 2011.[1] In the database used by the authors, there is a high missing rate (41%) of the primary outcome variable 'neurodevelopmental disorders' and high missing rates (between 31.9% and 49.4%) of the secondary outcome variables (eg, 'cerebral palsy at 5 years,' 'coordination disorders,' 'cognitive impairments,' and 'behavioral difficulties'). Rather than ignore the incomplete records the authors conducted statistical analyses after multiple imputation (MI) procedures were used, which has been regarded as an effective method to reduce nonresponse bias.

## Effect of Missing Data

The issue of missing data happens frequently in pediatric research. Missing data can be classified into 2 types: unit nonresponse and item nonresponse.[2] Unit nonresponse happens when respondents fail to answer a large portion of survey items whereas item nonresponse happens when respondents fail to answer moderate or small portions of survey items. The missing data found in the EPIPAGE 2 is considered item nonresponse since participants still answered maternal, obstetrical, and neonatal characteristics questions, even though they missed questions related to the outcome variables. The primary outcome variable 'neurodevelopmental disorders' had a missing rate of 41% and the secondary outcome variables 'cerebral palsy at 5 years,' 'coordination disorders,' 'cognitive impairments,' and 'behavioral difficulties' had missing rates of 31.9%, 49.4%, 40.7%, and 40.5%, respectively. Literature suggests that data sets containing 5% or less missing data are not likely to benefit substantially from MI,[3] while data sets with more than 10% missing data will tend to produce biased estimates if the missing data are not handled properly.[4] The extent of nonresponse bias is contingent upon both the response rate and the correlation between the underlying response probability (the probability of answering the survey question) and the outcome variable of interest.[5] With such high rates of missing data, EPIPAGE 2 may indicate a significant amount of nonresponse bias.

| | |
|---|---|
| EPIPAGE 2 | epidemiological study on small gestational ages |
| MAR | missing at random |
| MCAR | missing completely at random |
| MI | multiple imputation |
| NMAR | not missing at random |
| OR | odds ratio |

## Missing Mechanism

The reason for missing data, often referred to as the missing mechanism, can be classified into 3 types: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR).[2,6] MCAR refers to a type of missing data mechanism in which the probability of a data point being missing is unrelated to both observed and unobserved data. For example, suppose missing height measurements occurred because of a temporary malfunction in the clinic's electronic health record system. This malfunction randomly affected a small proportion of the height measurements, regardless of the children's actual heights or any other characteristics. This scenario illustrates MCAR in a pediatric context. MAR is a type of missing data mechanism in which the probability of a data point being missing depends only on observed data and not on the unobserved data. For example, in a study on toddlers' language development, researchers collect data on vocabulary growth by assessing spoken word counts over time. However, some toddlers may refuse to participate in 1 session, likely due to mood or behavior. These missing data, related to mood or behavior, can be accounted for in analysis, exemplifying data MAR. NMAR refers to a type of missing data mechanism in which the probability of data being missing is related to the unobserved data itself, even after accounting for the observed data. For example, imagine a study tracking the effectiveness of a new medication for children with asthma. During follow-up appointments, children who experience severe side effects from the medication are less likely to return for subsequent visits. Consequently, their data on symptom improvement becomes increasingly sparse over time compared with those without side effects. In this scenario, the missing data on symptom improvement are related to the severity of side effects, a factor not observed or easily accounted for in the analysis, making such an example of data NMAR. In many real world scenarios, it is difficult to ascertain whether data are MCAR or NMAR. MAR provides a middle ground, that is, often more plausible and pragmatic to assume, especially when there is no clear evidence to support either MCAR or NMAR. Overall, MAR offers a reasonable compromise between handling missing data appropriately and maintaining practicality in statistical analysis, which is why it is frequently used

From the Department of Biostatistics and Epidemiology, College of Public Health, The University of Oklahoma, Health Sciences Center, Oklahoma City, OK

0022-3476/$ - see front matter. 
https://doi.org/10.1016/j.jpeds.2024.114083

in practice. In their study, Salmon et al adopted an MAR assumption.[1] Their imputation model encompassed variables that could potentially predict nonresponse as well as those predicting outcomes. They incorporated a diverse array of predictors into the model to uphold the MAR assumption during imputation and to enhance the reliability of the imputed results.

## Handling Missing Data in Practice

In practice, methods for handling missing data include listwise deletion, pairwise deletion, single imputation, maximum likelihood, and MI.[7] Listwise deletion (cases with any missing values are entirely excluded) and pairwise deletion (available case analysis or case-wise deletion) may lead to biased results since they ignore cases with missing values in the analysis. Single Imputation is not recommended since it fails to capture the variability (uncertainty) due to imputation. Maximum likelihood can be used as an alternative method to the MI method. It is a statistical approach finding the best fitting model parameters that explain the observed data, even when some values are missing. MI, however, is one of the most effective tools for handling missing data in practice due to its flexibility for handling different data types, availability of computational software, and its ability to capture the variability (uncertainty) due to imputation. It can also ensure that different analysts produce consistent results after statistical analysis. Salmon et al used MI in their analysis due to the fact that multiple outcome variables were missing simultaneously.[1] In this case, MI can be used to build the joint model for multiple outcome variables and provide imputed data so that researchers can conduct different types of analysis with consistent results.

## MI

MI is a statistical technique used to handle missing data by generating multiple sets of plausible values for the missing data based on the observed data and statistical models.[8] The process involves creating multiple complete data sets, each containing different imputed values for the missing data, and then analyzing each data set separately to obtain parameter estimates. The results of the separate data sets are combined to produce overall estimates with their accompanying standard errors. MI is often preferred over single imputation methods because it provides more accurate, reliable, and less biased estimates. MI is widely used in various fields, including epidemiology, pediatrics, and clinical research, where missing data are common but simply ignoring the missing items or using ad hoc imputation methods may lead to biased results.

## Number of Imputations

The number of imputations in MI refers to how many separate data sets with imputed values are created to handle missing data. Choosing the appropriate number of imputations is important as it affects the accuracy and reliability of the imputation process. While there is no strict rule of thumb for determining the number of imputations, a common guideline is to use at least 20 imputations. Some researchers may choose more or fewer depending on the complexity of the data set and the amount of missing data. Increasing the number of imputations generally leads to more precise estimates but also requires more computational resources. Salmon et al generated 50 imputed data sets to ensure good efficiency and reliability of their estimates.[1]

## Rubin Rules

Rubin rules, named after Donald Rubin who is a pioneer in missing data statistical methodologies, are a set of principles used to combine the results obtained from analyzing multiple imputed data sets into overall parameter estimates and standard errors.[8] These rules allow for the appropriate incorporation of uncertainty introduced by the imputation process into the final analysis. By applying Rubin rules, researchers can obtain valid and efficient inferences from multiple imputed data sets while properly accounting for the uncertainty introduced by the missing data. These rules are widely used in practice to ensure the reliability of results in MI analyses. Rubin rules were used by Salmon et al to conduct statistical inference (eg, generating ORs and their 95% CIs) of the association between clinical chorioamnionitis and neurodevelopmental disorders at 5 years of age in children born preterm.[1] In this case, 50 ORs and 50 corresponding variances of ORs were first calculated separately based on 50 imputed data files. Then the average of 50 ORs was used as the final combined estimate for the OR. The final SE of the OR was calculated based on combining the within imputation variance (average of 50 variances of ORs) and between imputation variance (variance of 50 ORs).

## Evaluation of MI

Evaluating the performance of MI involves assessing the quality of the imputed values and the impact of the imputation on the results of subsequent analyses. Some common approaches to evaluate the performance of MI include: 1. comparison with complete data analysis (after deleting missing values); 2. assessment of imputation models (check model assumptions and fitting); or 3. sensitivity analysis (assess the robustness of the results to different model assumptions or imputation methods). By employing these approaches, researchers can assess the performance of MI methods and ensure the reliability of results obtained from analyzing imputed data sets. Assessment of imputation models in 2 above can be conducted by computing the model fitting statistics including Akaike information criterion, Bayesian information criterion, area under the receiver operating characteristic curve, and concordance index. Salmon et al conducted sensitivity analysis by comparing the MI results with complete data analysis.[1] The presence of clinical chorioamnionitis did not show an association with moderate-to-severe neurodevelopmental disorders in the complete cases analysis. Consequently, sensitivity analyses yielded results consistent with those of the main analysis by using MI.

## Interpretation of the Study Findings

In the paper by Salmon et al, a comprehensive set of predictors was used to construct the imputation model, serving to validate the assumption of MAR.[1] Following the MI process, analyses revealed that neither the crude associations nor the adjusted associations, considering gestational age at birth alone or with additional covariates such as household socioprofessional category, mother's country of birth, mother's level of education, smoking, multiple pregnancy, maternal obesity, and preterm premature rupture of membranes, yielded statistically significant results in relation to the primary and secondary outcome variables. These findings remained consistent with those obtained from the complete data analysis. ■

## CRediT authorship contribution statement

**Sixia Chen:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Michael P. Anderson:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

## Declaration of Competing Interest

S.C. and M.P.A. were supported by the Oklahoma Shared Clinical and Translational Resources (U54GM104938) with an Institutional Development Award from National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors declare no conflicts of interest.

Reprint requests: Sixia Chen, PhD, Department of Biostatistics and Epidemiology, College of Public Health, The University of Oklahoma, Health Sciences Center, Oklahoma City, OK. E-mail: sixia-chen@ouhsc.edu

## References

1. Salmon F, Kayem G, Maisonneuve E, Foix-L'Hélias L, Benhammou V, Kaminski M, et al. Clinical chorioamnionitis and neurodevelopment at 5 Years of age in children born preterm: the EPIPAGE-2 cohort study. J Pediatr 2024;267:113921.
2. Little RJ, Rubin DB. Statistical analysis with missing data, 793. Hoboken: John Wiley & Sons; 2019.
3. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999;8:3-15.
4. Bennett DA. How can I deal with missing data in my study? Aust N Z J Publ Health 2001;25:464-9.
5. Schouten B, Cobben F, Bethlehem J. Indicators for the representativeness of survey response. Surv Methodol 2009;35:101-13.
6. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7:147.
7. Newman DA. Missing data: Five practical guidelines. Organ Res Methods 2014;17:372-411.
8. Rubin DB. Multiple imputation. In: Flexible imputation of missing data. 2nd ed. Boca Raton: Chapman and Hall/CRC; 2018. p. 29-62.