# Development and Application of a Framework for Addressing New Challenges of Missing Data in Real-world Research

Zhaohui Su, PhD, Amy K. O'Sullivan, PhD, Kenneth Dwyer, MS, Janet Espirito, PharmD, Jessica K. Paulus, ScD

ontada

The US Oncology Network

MSR21

## Objectives

The rise in real-world data (RWD) sources poses challenges to handling missing data when combining data sources. Issues include data linkage, augmentation using natural language processing (NLP) and machine learning (ML) technologies, and reconciliation across diverse sources. We therefore developed a novel framework for handling missing data in a RW database that consists of multiple sources, such as electronic health records (EHR) and administrative claims.

## Results

### Table 1: Literature review of existing methods

| Amount of missing data | Missing data mechanism | Methods |
|---|---|---|
| < 5% | Any | Complete case, best-worst case analyses |
| 5 – 40% | Missing completely at random (MCAR), MAR | Multiple imputation, complete case and likelihood-based analyses, censoring methods |
| | Missing not at random (MNAR) | Pattern mixture and selection models |
| >40% | Any | Tipping point analyses, discuss limitations |

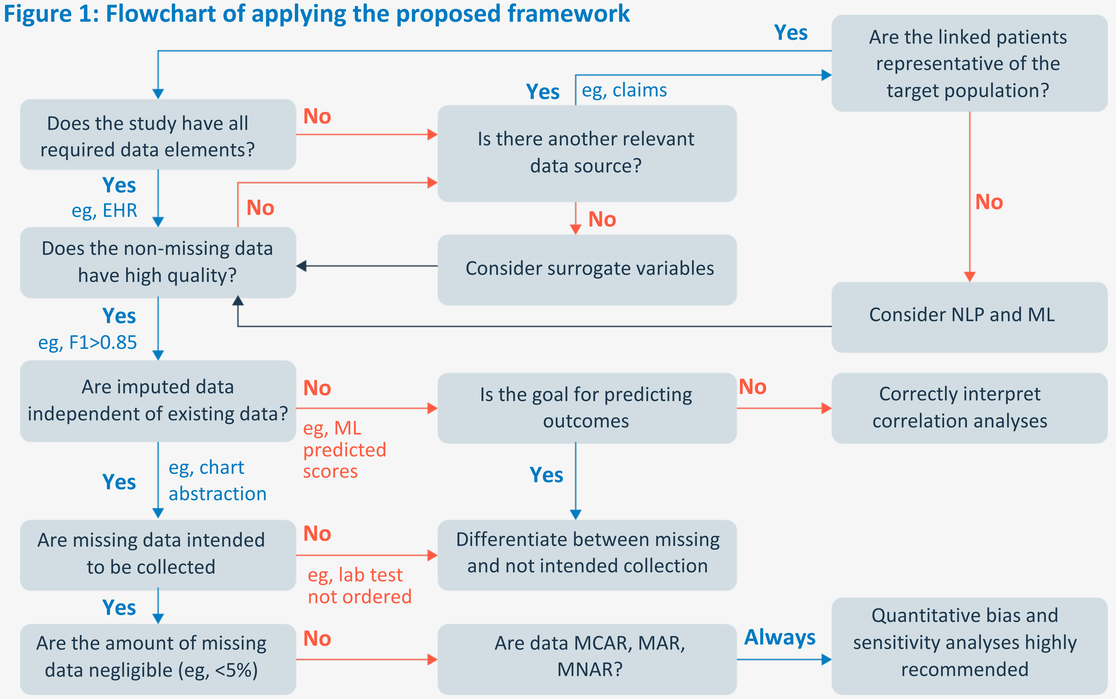
### Table 2: Proposed framework for handling missing data

| Domain | Description | Contributions to literature |
|---|---|---|
| Data relevance and representativeness | When multiple data sources are integrated, the representativeness of the overlapped patients or data should be examined, documented and reported. | Emphasizes the importance of representativeness when addressing missing data by linking data with external sources. |
| Data quality | Data quality of supplementary sources or from predictive technologies including NLP and ML should be evaluated following a fit-for-purpose framework. | With increasing data completeness by NLP and ML, data quality should remain a priority, and should be carefully evaluated. |
| Data correlation | Correlation among data elements and its impact on results should be assessed. | Results of correlation analysis should take into consideration how missing data are imputed. |
| Intended collection | Clinical expertise must be applied to differentiate between missing data, data not routinely available in RW standard of care, and data not intended to be collected. | RWD not intended to be collected should be checked and treated separately in the analysis, eg, with missing data indicator. |
| Quantitative bias and sensitivity analyses | Quantitative bias and sensitivity analyses are strongly encouraged. | Strongly recommends the use of quantitative bias and sensitivity analyses, eg, E-value |

## Methods

Published methods for handling missing data were reviewed (Table 1). Scientific experts representing population health and data science disciplines developed a new framework that builds upon existing rubrics to address unmet needs in reconciliation across multiple sources, including imputed values. The framework has 5 domains (Table 2) with a flowchart (Figure 1).

### Figure 1: Flowchart of applying the proposed framework




## Conclusion

- Existing analytic methods for addressing missing data focus on missing data volume and mechanism, and do not fully address the new challenges of missing data when using a multi-sourced real-world database.
- This novel framework provides an objective approach to maximizing completeness and describing validity concerns. It highlights the important considerations of data representativeness, quality and potential bias when handling missing data.
- The new framework supplements existing methods for handling missing data and will increase the quality of real-world evidence studies.