Imputing Missing Data in Observational Studies: What Methods Are Better?

Vincent McCarty[1], Paige Kostoulias[1], Neil R. Brett[2], Marielle Bassel[2], John Sampalis[1,2]

[1] Department of Surgery, McGill University, Montréal, Canada; [2] Thermo Fisher Scientific, Montreal, Canada

MSR211

Background

- Missing outcome data are common in longitudinal observational studies and may arise from loss to follow-up, irregular visit schedules, or incomplete data capture. If not appropriately addressed, missing data can introduce bias in estimates of disease progression and treatment effectiveness.
- The impact of missingness is particularly important for longitudinal disease activity measures, where patterns of change over time are central to interpretation.
- While several imputation methods are available, including single imputation, model-based methods, regression-based approaches, and multiple imputation, these rely on assumptions that may not hold in real-world settings.
- Despite their widespread use, there is limited comparative evidence on the extent to which these methods preserve true disease trajectories and minimize bias.

Objectives

- To compare the accuracy of four commonly used imputation methods (Last Observation Carried Forward [LOCF], Expectation-Maximization [EM], Linear Regression [REG], and Multiple Imputation [MI]) in replacing missing longitudinal disease activity data, and to assess which method introduces the least bias relative to original complete data (using rheumatoid arthritis [RA] as a case example).

Methods

- Data from 286 patients with RA in an observational study with complete longitudinal assessments for the following three disease activity indices were used:
  - Clinical Disease Activity Index (CDAI)
  - Simple Disease Activity Index (SDAI)
  - Disease Activity Score-28 using C-Reactive Protein (DAS28-CRP)
- Assessments were conducted at 0, 3, 6, 9, 12, and 18 months (1,716 total observations).
- Across all values except baseline, 10% of data points were deleted under a Missing Completely at Random (MCAR) mechanism. The missing data were then replaced using each of the four imputation methods.
- The four imputation methods compared were:

LOCF	Replaces missing values with the most recent non-missing observation for the same patient.
EM	Iteratively estimates parameters using maximum likelihood, assuming a multivariate normal distribution.
REG	Predicts missing values from observed data using a linear regression model.
MI	Generates multiple plausible datasets, analyzes each, and pools results using Rubin's rules.

- Imputed datasets for each disease activity index were compared with the original complete dataset using one-sample t-tests and relative percentage differences.

Results

Table 1. One-Sample t-test Results

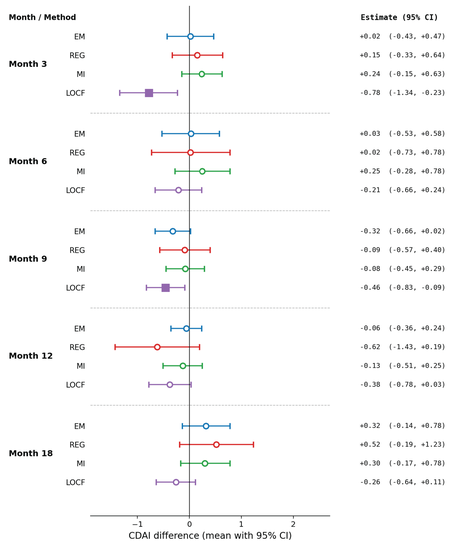
Outcome / Method	Mean Difference	t	df	p (2-tailed)	Relative Difference (%)
CDAI					
EM	0.097	1.023	1715	0.307	0.59%
REG	0.102	0.733	1715	0.464	0.62%
MI	0.121	0.048	1715	0.962	0.73%
LOCF	**-0.347**	**-2.003**	**1714**	**0.045**	**-2.10%**
SDAI					
EM	-0.079	-0.885	1715	0.376	-0.45%
REG	-0.064	-0.516	1715	0.606	-0.36%
MI	-0.030	-0.318	1715	0.751	-0.17%
LOCF	**-0.564**	**-4.216**	**1704**	**<0.001**	**-3.19%**
DAS28-CRP					
EM	-0.010	-1.070	1715	0.285	-0.32%
REG	0.005	0.397	1715	0.691	0.16%
MI	-0.000	-0.033	1715	0.974	-0.01%
LOCF	**-0.023**	**-5.243**	**1706**	**<0.001**	**-0.72%**

Mean Difference = Imputed Mean – Original Mean. Relative Difference = Mean Difference / Original Mean × 100. Bolded rows have *P*<0.05. df varies for LOCF due to residual missing values where no prior observed value was available for carry-forward.

Abbreviations: CDAI = Clinical Disease Activity Index; DAS28-CRP = Disease Activity Score-28 using C-Reactive Protein; EM = Expectation-Maximization; LOCF = Last Observation Carried Forward; MI = Multiple Imputation; REG = Linear Regression; SDAI = Simple Disease Activity Index

CDAI (Figure 1):

Figure 1. Forest Plot: CDAI Difference by Imputation Method and Month

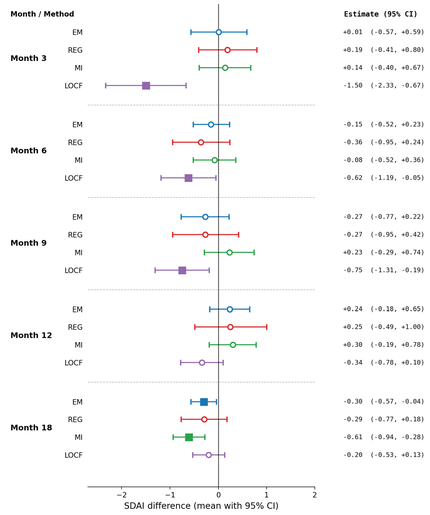


Abbreviations: CDAI = Clinical Disease Activity Index; EM = Expectation-Maximization; LOCF = Last Observation Carried Forward; MI = Multiple Imputation; REG = Linear Regression

- EM, REG, and MI demonstrated minimal bias across timepoints, while LOCF-imputed means were significantly different from original data at months 3 (–0.78; 95% CI: –1.34, –0.23) and 9 (–0.46; 95% CI: –0.83, –0.09)

SDAI (Figure 2):

Figure 2. Forest Plot: SDAI Difference by Imputation Method and Month

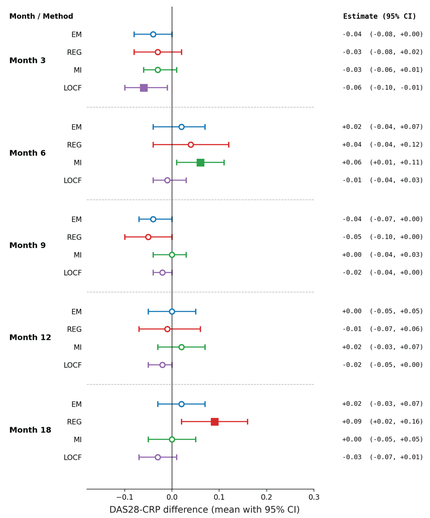


Abbreviations: EM = Expectation-Maximization; LOCF = Last Observation Carried Forward; MI = Multiple Imputation; REG = Linear Regression; SDAI = Simple Disease Activity Index

- REG demonstrated minimal bias across timepoints. LOCF-imputed means were significantly different from original data at months 3 (–1.50; 95% CI: –2.33, –0.67), 6 (–0.62; 95% CI: –1.19, –0.05), and 9 (–0.75; 95% CI: –1.31, –0.19), while EM and MI-imputed means were significantly different from original data at month 18 (EM: –0.30, 95% CI: –0.57, –0.04; MI: –0.61, 95% CI: –0.94, –0.28).

DAS28-CRP (Figure 3):

Figure 3. Forest Plot: DAS28-CRP Difference by Imputation Method and Month



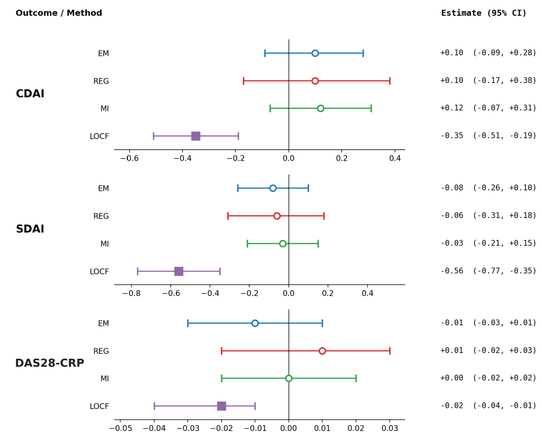
Abbreviations: DAS28-CRP = Disease Activity Score-28 using C-Reactive Protein; EM = Expectation-Maximization; LOCF = Last Observation Carried Forward; MI = Multiple Imputation; REG = Linear Regression

- All methods demonstrated minimal bias across the timepoints and closely aligned with the original data; differences between imputation methods were small.

Results (continued)

Overview (Figure 4):

Figure 4. Forest Plot: Outcome Differences by Imputation Method (95% CI)



Abbreviations: CDAI = Clinical Disease Activity Index; DAS28-CRP = Disease Activity Score-28 using C-Reactive Protein; EM = Expectation-Maximization; LOCF = Last Observation Carried Forward; MI = Multiple Imputation; REG = Linear Regression; SDAI = Simple Disease Activity Index

- EM, REG, and MI aligned with the original data across disease activity indices, while LOCF-imputed means were significantly different, with the largest magnitude of difference observed for CDAI and SDAI.

Discussion

- Overall, LOCF-imputed means were significantly different from original data for all three outcomes: CDAI (P=0.045), SDAI (P<0.001), and DAS28-CRP (P<0.001). All other methods (EM, REG, MI) were more consistent with the original data.
- The overall relative difference between imputed and original data ranged from 0.72% to 3.19% for LOCF and 0.01% to 0.73% for the other methods.
- MI produced the smallest differences for two of three indices, supporting its use as the preferred imputation method.
- Limitations include the use of simulated missingness (10% MCAR), which may not reflect real-world missing data patterns, and the analysis was limited to RA disease activity measures.
- Future work should examine performance under missing at random and missing not at random mechanisms and in other therapeutic areas.

Conclusions

- In this analysis, EM, REG, and MI-based imputation produced more reliable results than LOCF across all three disease activity indices (CDAI, SDAI, DAS28-CRP).
- Based on the t-tests and P-values, MI imputations were closest to the original data.
- These findings support the potential for additional exploration of the use of MI for handling missing outcome data in real-world longitudinal studies, and caution against the routine use of LOCF.

Disclosures

Funding provided by Thermo Fisher Scientific. **NRB**, **MB**, and **JS** are employees of PPD™ Observational Studies, Thermo Fisher Scientific at the time this study was conducted. **VM** and **PK** have no conflicts to disclose.