Multiple Statistical Imputation of Missing Data in a Real World Clinical Cohort Increased Sample Size and Power with Minimal Deviations Enhancing the Clinical Research Potential of the Observational Dataset

Authors: Tatangelo, M., Li, X., Cesta, A., Movahedi, M., Bombardier, C., Tomlinson, G.

Objectives: Real-world observational studies occur amid the context of day-to-day clinical care where missing data including data not collected or patients lost at follow-up are unique when compared to clinical trials. Clinical trials use large budgets and strict control of clinical practice to avoid missing data. In order to continue to reflect the realities of clinical care, imposing strict guidelines on care to reduce missing data is not preferable. We sought to use the current best practices in clinical observational research, multiple statistical imputation (MSI), to assess the difference in data quality when MSI was used compared to when data were simply left missing.

Methods: Core variables were selected based on their frequency of use by researchers querying data from the Ontario Best Practices Research Initiative, an observational clinical cohort of adult patients with incident or prevalent active rheumatoid arthritis (RA) in Ontario (n=3075) with n=56 rheumatologists contributing patients. The variables selected were patient demographics, RA disease activity (DAS28, CDAI), and HAQ score. MSI works by replacing the missing values for each patient with a plausible substitute based on the complete data from all patients that are similar to the patient with the missing data value. In order to determine if MSI improved data quality we used statistical models (Generalized Linear Models) to compare an identical analysis in the OBRI dataset including missing data, then again in the OBRI dataset with no missing data due to MSI.

Results: A total of 3075 patients were included in the study, with 47,703 visits. Results using the original data with missing values compared to the complete data after imputation were almost identical Db=0.00130.02, with a reduced standard error 0.150.08.
Clinically, this means the models had a small difference in DAS28 of less than 1% and a decrease in standard deviation of 41% for clinical research using MSI data instead of datasets containing missing data.

Conclusion: Because the differences in models were small, results from the MSI dataset were functionally identical yet had narrower confidence intervals. Therefore, the use of MSI in the OBRI clinical cohort increased sample size and power while minimizing standard deviation and confidence intervals enhancing the clinical research potential of the observational dataset. MSI represents the most methodologically rigorous approach to missing data thus minimizing error in clinical results. Future analyses from querying the OBRI database will rely on MSI.