Missing data imputation in longitudinal trial of endometrial cancer patients

Missing data imputation in longitudinal trial of endometrial cancer patients

Rosato R, Pagano E, Piovano E, Fuso L, Tripodi E, Mitidieri M, Ceccarelli M, Zola P, Di Cuonzo D

Aims

Missing data is a common problem in health-related quality-of-life studies, especially with longitudinal study design.
Several techniques for the multiple imputation (MI) for missing data have been developed and validated on simulated data.
This study aims to investigate the missing data generating process, define models for MI and validate them on real data.

Methods

Data of all the patients recruited from September 2008 to May 2014 in an ongoing randomized clinical trial between two follow up regimens in endometrial cancer surgically treated patients were used.
Participants were followed since the end of primary treatment with repeated clinical evaluation and quality of life (QoL) assessments (baseline, 6, 12, 24,36,48, 60 months) with the Short Form 12-item survey (SF-12).
QoL missing data did not satisfy the assumption of multivariate normality.
We used two different techniques to impute non-monotonic data: Markov-chain Monte Carlo (MCMC) and fully conditional specification (FCS).
In order to impute missing data using FCS method, we specified a separate regression model for each variable with missing.
Validation of the quality of the process of imputation was performed on data collected from May 2014 to November 2015 for the patients still in follow-up.
For each MI approach, we assessed bias (average difference between estimate and real data) and coverage of the estimate 95% confidence interval for each coefficient estimate to determine which procedure is most appropriate.

Results

In the study period, 822 patients were enrolled.
Using the two different methods of imputation similar results were obtained on all the measurement points.
We found no evidence that MCMC method performed less well, despite the unrealistic assumption of multivariate normal distribution.
Both the methods show a good performance.
For example, the absolute bias for Mental Component varied between 0.03 and 1.6.
Performance was slightly reduced in the long term data.

Conclusions

To the best of our knowledge, this study is the first to use real data to validate the two techniques of imputation on longitudinal records with a large number of measurement points and a large proportion of missing data.
Our results are in line with previous researches based on simulated data.