ENHANCING CANCER REGISTRATION DATASETS: COMPARISON OF ALGORITHMS FOR MULTIPLE IMPUTATION OF MISSING VALUES

MATTHIAS LOREZ[1] AND ANDREA BORDONI[2]

[1] National Institute for Cancer Epidemiology and Registration (NICER), Zurich, Switzerland; [2]Ticino Cancer Registry, Institute of Pathology, Locarno, Switzerland

BACKGROUND

Missing data constrain the value of population-based cancer registries in cancer control programs. Multiple Imputation (MI) reduces bias in statistical inference from incomplete datasets, as compared to simple complete case (CC) analysis. We compared two algorithms for MI: Chained-Equations (MICE) and Expectation-Maximization applied to bootstrapped data (EMB).

METHODS

We simulated 30% and 60% univariate missingness in tumor stage in breast cancer registered 1996-2005, using 6 different missingness mechanisms. Cox models were fitted to complete, incomplete and imputed datasets. Analysis endpoint was the stage-dependent hazard ratio. Regressions included stage as ordinal (A1) or continuous (A2).

RESULTS

CC analysis generated large biases only with 60% missingness. Estimates after MICE were always close to complete data. EMB introduced large biases with regression model A1 (stage violates the normality assumption) but not A2.

CONCLUSIONS

MI is superior to CC analysis; provided that analyses are robust to deviations from distributional assumptions in the imputation algorithm. We favoured MICE because tailored distributions are used for imputed variables. The advantage of EMB is significantly faster processing.