Enhancing cancer registration datasets: Comparison of algorithms for Multiple Imputation of missing values

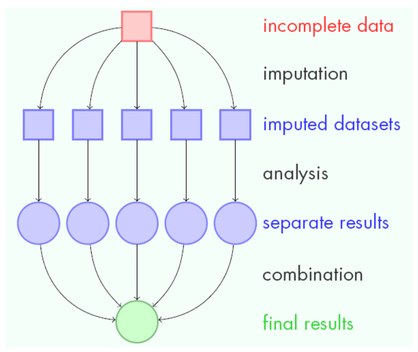
Matthias Lorez[1] and Andrea Bordoni[2]

[1]Foundation National Institute for Cancer Epidemiology and Registration (NICER), Zürich, Switzerland. [ml@nicer.org]
[2]Ticino Cancer Registry, Institute of Pathology, Locarno, Switzerland. [andrea.bordoni@ti.ch]

1. The Problem

Datasets in population-based cancer registration often contain missing values. The most common approach to this problem is to restrict analysis to those cases without missing values (Complete Case analysis, CC). Depending on the amount of missing information, this approach will reduce the precision of estimates considerably and may introduce bias if information is not missing completely at random.

2. Background: Algorithms for Multiple Imputation



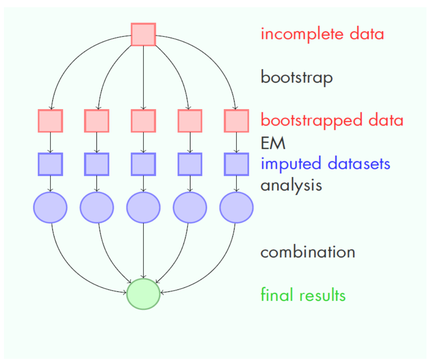
Multiple Imputation (MI) aims at reduced tendency for bias and higher efficiency in statistical inference from incomplete datasets. MI is based on the assupmtion that values are missing due to 'ignorable' mechanisms (i.e. not depending on unobserved data). Thus, missing data can in principle be inferred from the available data.
In short, each missing value is filled in with multiple random draws from the conditional predictive distribution of missing values. Thus, m imputed datasets (usually 3 to 10) are generated and each analysed using standard methods. The m separate results are then combined to obtain a single final result.

We compared two algorithms for MI:

MICE (Multiple Imputation by Chained Equations) and

EMB (Expectation-Maximization to Bootstrapped data)

MICE is implemented by the statistical program ICE v1.6.7 (Ref.1,2) in Stata™. It initiates with random draws from observed data and updates these values by conditioning on predicting variables with separate variable-specific regression equations, not limited to the multivariate normal assumption. The algorithm iterates sequentially over these conditional densities similar to Gibbs sampling until convergence is reached. MICE thus avoids the formulation of a single joint distribution model. These steps are repeated m-times to generate m imputed datasets.



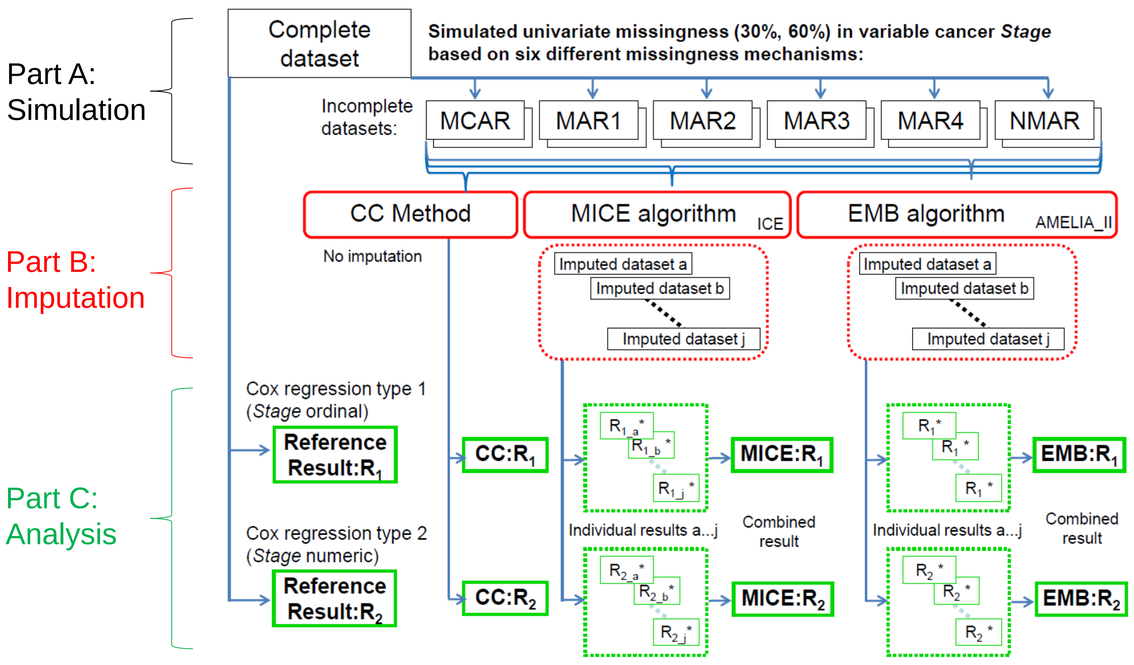
The EMB-Algorithm is implemented by the software AMELIA_II v1.16 (Ref.3) in the R language.
This approach first bootstraps the unimputed data m-times to recover population variance. The distribution parameters of the m incomplete datasets are estimated from the observed data by Expectation-Maximization (EM) assuming multivariate normality for observed and unobserved values. Single draws conditioning on observed values and the m derived distribution parameters replace missing values and generate m imputed datasets.

3. Steps in the analytical procedure

Part A: The reference dataset of 1819 female patients diagnosed with 1st primary breast cancer in the canton of Ticino (Switzerland) was without missing values. We simulated univariate missingness in the variable cancer *Stage* due to various missingness mechanisms: ‚completely at random' (MCAR), ‚missing at random' (MAR) depending on the variable *Lifestatus* (MAR1), *Follow-up duration* (MAR2), *Age-at-diagnosis* (MAR3) or *Diagnosis-year* (MAR4) and ‚not missing at random' depending on *Stage* itself (NMAR). In addition, levels of missingness were simulated as moderate (30%) or high (60%).

Part B: The 12 incomplete datasets were either analysed directly with the CC method or after application of multiple imputation. All imputation models included the following variables: *Follow-up duration* (Box-Cox transformed), *Age-at-diagnosis*, *Diagnosis-year*, cancer *Stage* (ordinal or numeric), *Tumour grade* and *Tumour size* (log transformed).

Part C: We chose the hazard ratio of patients with cancer *Stage* 3 versus *Stage* 1 as principal outcome of interest. Two different Cox regression models with cancer *Stage* and *Age-at-diagnosis* as covariates were applied. The 1st model included *Stage* as ordinal, the 2nd as numeric variable.
The approximate bias of the three missing-data handling methods (CC, MICE, EMB) was assessed by the difference in hazard ratios to the ‚true' hazard ratio observed in the complete dataset.



4. Reference Results: Complete dataset

Cox regression type 1 (*Stage* ordinal):

```
No. of subjects =        1819                    Number of obs   =      1819
No. of failures =         329
Time at risk    =  9181.78234
                                                 LR chi2(3)      =    287.97
Log likelihood  =  -2138.8313                    Prob > chi2     =    0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Stage_2	1.848067	.2626193	4.32	0.000	1.398807	2.441617
Stage_3	→ 4.72014	.7336424	9.98	0.000	3.480601	6.401113
Age_diagnosis	1.059022	.0047533	12.78	0.000	1.049746	1.068379

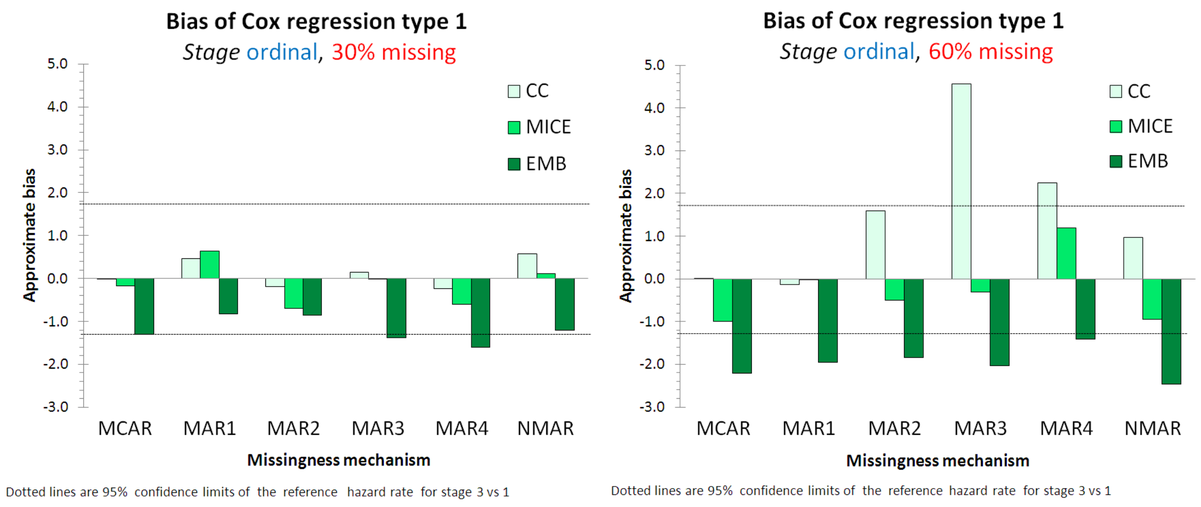
Cox regression type 2 (*Stage* numeric):

```
No. of subjects =        1819                    Number of obs   =      1819
No. of failures =         329
Time at risk    =  9181.78234
                                                 LR chi2(2)      =    285.86
Log likelihood  =  -2139.8893                    Prob > chi2     =    0.0000
```

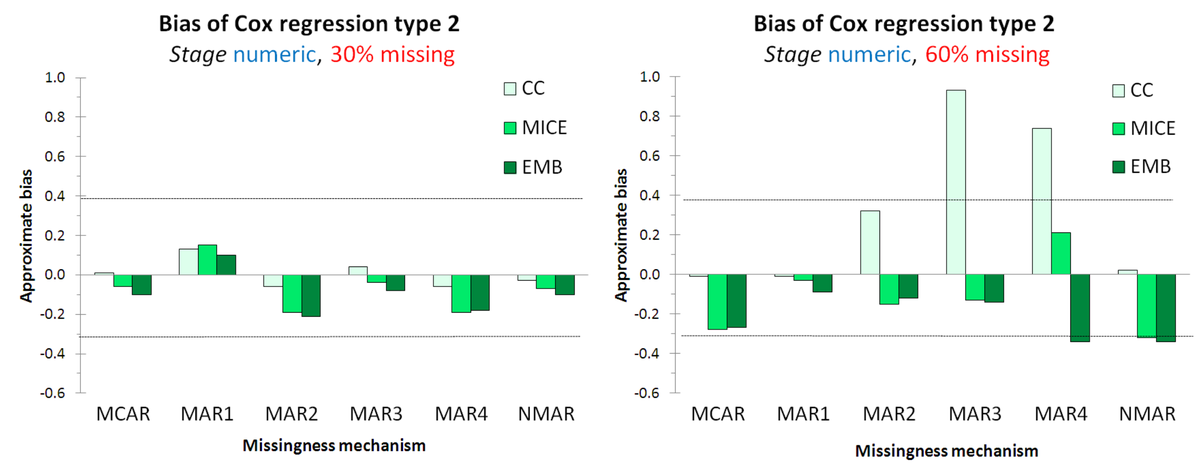
_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Stage	→ 2.200195	.1769226	9.81	0.000	1.879378	2.575777
Age_diagnosis	1.058792	.0047517	12.73	0.000	1.04952	1.068146

Using *Stage* as numeric variable implies that the influence of *Stage* would be entirely linear. This assumption seems justified since the hazard ratio of ordinal *Stage* 3 vs 1 of 4.72 is similar to twice the unit increase in hazard ratio for numeric *Stage* of 2.20.

5. Comparison of Results by CC, MICE and EMB



Dotted lines are 95% confidence limits of the reference hazard rate for stage 3 vs 1



Dotted lines are 95% confidence limits for the unit increase in the reference hazard rate.

The simple CC approach to missing data gave satisfactory results only if missingness was restricted to 30% (left panels). This was true for all mechanisms of missingness and for both regression models. The biases for MI by MICE were comparable to CC at 30% but substantially smaller at 60% missingness (right panles). MI by EMB introduced large biases in combination with regression type 1 (*Stage* ordinal). EMB-biases were much improved if *Stage* was modeled as a numeric variable in regression type 2. The missingness mechanisms also played a role: MI was not superior to CC in the case of NMAR because it violated the assumption of ignorable missingness.

6. Learning experiences

Approach to missing data	Strengths	Weeknesses
Complete Case Analysis	• Simple, fast • May be acceptable if ≤ 30% missingness	• Inefficient estimation (waste of data) • Possibility of bias if ≥ 30% missingness
MICE-Algorithm	• Flexible assumptions • Robust for different missingness mechanisms • Robust for different types of analysis	• Slow processing • Lack of theoretical basis (justification rests on empirical studies)
EMB-Algorithm	• Fast processing (designed for large datasets with multivariate missingness) • Time-series data imputation features available • Prior information for missing values can be incorporated	• Requires transformation of variables with skewed distribution • Possibility of enlarged biases due to multivariate normal assumption: must be checked for intended type of analysis

7. Conclusions

We consider MI as being superior to CC analysis in the presence of high levels of missing data, under the condition that the robustness of the intended analyses with respect to simplifying assumptions in the imputation algorithm has been investigated. The MICE algorithm offers the possibility to specify separate distributions for each imputed variable. The price for this versatility is prolonged processing time and the choice of imputation algorithm thus also depends on the size of the dataset from which inferences are intended.

References

1. Van Buuren et al. (1999). *Statistics in Medicine* 18, 681ff. 2. Royston (2000). *The Stata Journal* 9, 466ff.
3. Honaker and King (2010). *American Journal of Political Science* 54, 561ff.

Acknowledgments

We thank Prof. Jürg Hüsler, Institute of Mathematical Statistics and Actuarial Science, University of Berne, Switzerland for advice and support.