Using multiple imputation methods to estimate relative risks in small EPIC lung cancer subsets

Altenburg H.P.[1,2], Agudo A.[1], Berrino F.[1], Boshuizen H.C.[1], Bueno-de-Mesquita H.B.[1], Janzon L.[1], Le Marchand L.[1], Linseisen J.[1,2], Lukanova A.[1], Rasmuson T.[1], Vineis P.[1], Riboli E.[1], Miller A.[1,2]

(for the EPIC working group on lung cancer)

[1]EPIC. [2]German Cancer Research Center, Dept. of Clinical Epidemiology, Heidelberg, Germany.

Purpose

Although most of the relevant variables in the lung cancer subcohort contain only small proportions of missing values, subgroups are possible with higher proportions of missing data. Conventional methods for missing data, such as listwise deletion (as in most statistical software packages) or regression imputation may be sensitive to problems such as:

- Inefficient use of the available information, leading to low power and type II errors.
- Biased estimates of standard errors, leading to incorrect *P* values.
- Biased parameter estimates, due to failure to adjust for selectivity in missing data.

Moreover, listwise deletion of observations is inefficient, because it can dramatically decrease the number of cancer cases in the subgroup. We have similar problems when using methods such as pairwise deletion (available cases), single deterministic imputation, single random imputation or dummy variable adjustment.

More accurate and reliable results can be obtained using maximum likelihood or multiple imputation procedures. In case of data that are missing at random, both methods deliver approximately unbiased as well as efficient parameter estimates and standard errors. Well-developed maximum likelihood methods are *not* available for logistic regression or Cox regression. Therefore the objective of the paper was the use of multiple imputation methods described in Rubin (1987, 1996) or Schafer (1997) to estimate relative risks in subgroups of the lung cancer cohort.

Methods

Multiple imputation inference is performed in three distinct steps:

1. Missing values are represented by a random sample of size five from an appropriate imputation model incorporating random variation. This leads to five complete data sets to be analysed. First of all, one has to choose an appropriate set of variables, e.g. all variables that should be in the intended model, including the dependent variables and other characteristics that may be associated with variables that have missing data. If necessary, one has to transform the variables to achieve approximate normality. Then, running the Expectation-Maximum-Likelihood algorithm provides maximum likelihood estimates of means and the covariance matrix (used as starting values for the imputation process). Next, the actual data augmentation process is run to generate the multiple data sets with imputed values. Finally, the transformed data values have to be transformed back and imputed values of discrete variables have to be rounded.
2. The desired analysis is performed on each of these five data sets using standard software packages, meaning we analyse the data using complete-data methods.

3. The results of the five parameter estimates from each dataset are then combined. We average the parameter estimates across the five samples to obtain a single estimate. The corresponding standard errors are computed by combining the variation within and between the five samples.

The advantage of multiple imputation is that the random error in the imputation process yields approximately unbiased estimates of all parameters. The repeated imputation process gives us good estimates of the standard errors.

Two imputation mechanisms were used:

1. The propensity score and the Markov-Chain-Monte-Carlo (MCMC) mechanism. The propensity score (a nonparametric approach) is the conditional probability of assignment to a particular result given a vector of observed covariables. A propensity score is generated for each variable with missing values, indicating the probability of the observation being missing. The observations are then grouped based on this score and an approximate bootstrap imputation will be applied to each group.
2. The MCMC mechanism (appropriate for arbitrary missing patterns) constructs a Markov chain long enough for the distribution of the elements to stabilize to a common stationary distribution. By repeatedly simulating steps of the chain, it simulates draws from the underlying distribution of the data. The data augmentation process is a Bayesian inference consisting of an imputation step and a posterior step where the information on the unknown parameters is expressed in the form of a posterior distribution. Both steps have to be iterated long enough to get reliable results, thus to reach a stationary distribution and then to simulate an approximately independent draw of the missing values.

Results

Using multiple imputation methods one can find that the coefficient variance estimates can be up to 30% smaller than when using listwise deletion. But the propensity score method can lead to biased (larger) estimates of relative risks, whereas the MCMC method works well.

Conclusion

Use MCMC and include as many variables in the model as necessary. The MCMC method is appropriate for arbitrarily missing data, as is the case in EPIC subcohorts. The relative efficiency varies between 90% and 99% depending on the fraction of missing information on the corresponding parameter.

References

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys.* New York, John Wiley & Sons, Inc.

Rubin, D.B. (1996) Multiple imputation after 18+ years. *J. Am. Stat. Asso.*, **91**, 473–489

Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data.* New York, Chapman and Hall