Journal of Biometrics & Biostatistics
ISSN: 2155-6180
Yu et al., J Biom Biostat 2015, 6:3
DOI: 10.4172/2155-6180.1000243
Research Article
Open Access

Adaptive Robust Estimators to Handle Missing Values in Estimating Tumor Stage Distributions in Population-Based Cancer Registration

Qingzhao Yu[1]*, Han Zhu[1] and Xiaocheng Wu[2]

[1]*Biostatistics Program, School of Public Health, Louisiana State University Health Sciences Center, USA*
[2]*Louisiana Tumor Registry, Louisiana State University Health Sciences Center, USA*

Abstract

Accurate cancer stage at diagnosis is essential not only for assessing quality of care and associated prognosis but also for monitoring trends in cancer stages and for assessing effectiveness of early detection interventions. Because the cancer stage is associated with many factors that are not under control of cancer registries, it is infeasible to completely record stages in all cases from registry database. It is necessary to reduce the bias in stage analysis induced by unknown stage cases through statistical adjustment. In this paper, we propose a new adaptive robust method that estimates the distribution of unknown stage cases using both essential and nonessential predictors of cancer stage. Multiple additive regression trees were used to assess the association of explanatory variables (including patient demographics, tumor characteristics, and treatment) with unknown stage. The 2004-2009 incidence data on invasive lung cancer from 38 population-based cancer registries that met NAACCR's high data quality criteria were used to estimate the population stage distribution of lung cancer over the years. Multiple artificial incomplete datasets with unknown stages and predictors were created from the complete datasets, with varying missing data mechanism and different proportions of missingness. The simulated datasets were used to test the efficiency of the proposed method in estimating population stage distribution. In general, the proposed method is more efficient in terms of estimation accuracy and time consumption, compared with the traditional methods such as multiple imputation method and weighting method.

Keywords: Cancer stage; Robust estimator; Inverse probability weights; Missing data; MART; Multiple imputation

Introduction

Accurate data on cancer stage at diagnosis is essential for evaluating quality of care and associated prognosis, for monitoring trends in cancer stages, for assessing effectiveness of early detection interventions, and for measuring disparities in access to cancer care. Unknown stage cases might bring in a bias to data analysis especially when the stages are not missing at random. Using the 2004-2007 incidence data from population-based cancer registries, the Data Assessment Work Group of the NAACCR's Data Use and Research Committee found that the proportion of unknown stage cases varied substantially across cancer sites and cancer registries: 1.0%-13.7% for breast cancer, 0.6%-18.1% for prostate cancer, 2.4%-18.8% for colorectal cancer, 2.1%-18.7% for lung cancer, and 0.5%-14.4% for cervix cancer. Many factors are associated with the variation in proportions of unknown stage. Using multiple additive regression trees (MART) guided linear mixed effects model, Fan et al. [1] found that unspecific histology, non-microscopic confirmation, non-hospital reporting source and certain demographic characteristics were associated with high proportion of unknown stage. Since the missingness of cancer stage is associated with many factors that are not controllable by cancer registries, it is impossible to completely record stage information for cases in registry database. It is necessary to minimize the bias induced by unknown stage cases using statistical adjustment. The goal of this manuscript is to accurately estimate the population distribution of cancer stage based on the cancer registry databases. In this paper, we use the SEER summary stage 2000 to categorize the cancer stages (Young et al., 2001). In situ tumors fulfill all microscopic criteria for malignancy except invasion of the basement membrane of the organ. We consider only non-in situ cancer cases which leaves the stages localized, regional or distant. A "localized" tumor is confined to the organ of origin without extension beyond the primary organ. "Regional extension" of tumor can be by direct extension to adjacent organs or structures or by spread to regional lymph nodes. If the cancer has spread to parts of the body remote from the primary tumor, it is recorded as "distant" stage. For a single observation, cancer stage and/or some associated factors may be missing.

According to the general missing data mechanisms [2], stage information could be missing in the following patterns: 1) missing completely at random (MCAR), in which the stage distribution of unknown stage cases is the same as that of known stage cases; 2) missing at random (MAR), where the probability of missing a stage depends directly upon variables other than stage information; and 3) missing non-ignorable (MNI), where the probability of missing a stage depends not only on other variables but also on the stage itself even after controlling for all predictors, for example, when later stages are more likely to miss than early stages.

Statistical methods are available to deal with the MCAR and MAR missing problems. There are mainly three sets of methods: ad-hoc method [3], multiple imputations [4], and weighting method [5]. With the ad-hoc method, any observation with unknown stage is deleted from further data analysis. Multiple imputations are to predict/estimate missing values in the data set. The estimated values replace the missing data, and then the complete data sets are used for further analysis. There are two major approaches to impute multivariate data: joint modeling

***Corresponding author:** Qingzhao Yu, Associate professor, Biostatistics Program, School of Public Health, Louisiana State University Health Sciences Center, USA, Tel: (504) 568-6086; E-mail: qyu@lsuhsc.edu

Received July 28, 2015; **Accepted** August 07, 2015; **Published** August 14, 2015

Citation: Yu Q, Zhu H, Wu X (2015) Adaptive Robust Estimators to Handle Missing Values in Estimating Tumor Stage Distributions in Population-Based Cancer Registration. J Biom Biostat 6: 243. doi:10.4172/2155-6180.1000243

(JM) [6] and multivariate imputation by chained equations (MICE) [8]. JM method assigns a multivariate distribution for the missing data, and then imputes the missing values from the conditional distributions by Markov Chain Monte Carlo (MCMC) [7] method. In MICE algorithm, a multivariate imputation model is specified separately for each variable treating all other variables as predictors in the model. Stage by stage, one variable is selected, of which all missing values are imputed using the imputation model, then in turn the imputed variables are used to estimate the missing data in the next selected variable. This process repeats until certain convergence criteria is met [8]. To deal with the uncertainties in the multiple imputation process, several sets of imputations have to be generated, on each of which further analysis is implemented and the results are consolidated for the final inferences [8,9]. Weighting method excludes unknown stage cases from analysis. However, known stage cases are weighted to represent all cases in the population. Weights of the known cases are calculated through a weight model, where the response variable is the binary indicator of missing stage, and all the related risk factors are used as explanatory variables [10]. Typically, the weight of an observation is the inverse of the probability that the stage is known [5].

It is well known that the ad-hoc approach is simple and efficient when missing pattern is MCAR or the rate of missing is ignorable. Otherwise, the ad-hoc method could produce bias in estimating population stage distribution [11]. Multiple imputations could yield a valid estimation for stage distribution when the MCAR or MAR missing patterns applies [12], but multiple imputations can be very complicated and the stage wise imputations often bring in more uncertainties. Weighting method is generally simpler than the imputation method. However, the weighting method is very unstable when the weight model produces very large weights on a few observations [13]. The essence of the imputation method is that the predictive model has to be correct [8]. Similarly, it is important to have accurate weight model to apply the weighting method. There are methods that combine the weighting and multiple imputation methods so that as long as at least one of the predictive and the weight models is correct, the estimation is reliable. The method is called double robust (DR) procedure [14,15].

The purpose of this study is to apply an adaptive robust (AR) estimator to handle missing values in estimating cancer stage distributions. The rest of the paper is organized as following: Section 2 introduces the special characteristics in estimating stage distributions and then proposes the AR method. We apply the method to estimate the stage distribution of lung cancer patients diagnosed between 2004 and 2009 in Section 3. In Section 4, we conduct a simulation study based on the lung cancer data to assess the accuracy and efficiency of the proposed method in comparison with traditional methods. Finally, we discuss the strength and limitation of the proposed method and the direction of future research in Section 5.

Adaptive Robust Estimators

We propose an adaptive robust process to estimate the population cancer stage distributions. Dataset on cancer stage distribution has many special characteristics [16,17]. First, since we are dealing with population-based data, the number of cases is normally large. For example, the estimated number of new cancer cases per year ranges from 2, 460 other oral cavity cancer to 238, 590 prostate cancer (the American Cancer Society website http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2013/index). The big sample size results in high powers of finding true predicting/weighting models, even for nonparametric models that do not make special assumptions in model building. In such situation, nonparametric models are safer than parametric predicting/weighting models as fewer assumptions, such as linear relationship and normal distribution, are made. However, the predicting model for multiple imputations can be very complicated and it requests long computing time, especially that MI requires many repetitions to account for the randomness of the process [18]. Weight method is more convenient, but since the nonparametric model tend to over fit the data, some estimated probability of non-missing may be very small. Consequently, some known cases are assigned with large weights, which results in unstable estimation of stage distributions [19].

Another special character of stage information collected by the population-based cancer registration is that the stage levels are not directly observed, but are derived from essential tumor characteristics such as tumor extension, lymph nodes status, metastasis condition and other factors related to specific cancer sites. Therefore, stage may be coded as unknown if any one of the essential characteristics is missing, even when other characteristics may provide useful information on staging. Moreover, the stage distributions vary at different levels and different missing patterns of the risk factors, especially the essential tumor characteristic factors. In this paper, we propose a novel AR strategy to estimate the stage distribution of cancer population. The method can also be used in other fields to estimate population distributions.

The algorithm for AR estimation of population distributions

As discussed in Section 1, to estimate population stage distribution, ad-hoc method is most efficient when the missing of stage is completely at random or if the proportion of missing in stage is ignorable (e.g., smaller than 0.5%). If the missing of stages is not completely at random and the proportion of missing is significant, the weighting method is efficient when the weighting model is accurate and when there is no very large weight assigned to observed cases. MI is useful when the weighting method is not stable. The main idea of the proposed AR method is to split the covariate space into sub regions based on important factors related with stage, and then in each sub region, adaptively choose an effective method for distribution estimation. The method is robust since if any of the weighting or predictive models are not accurate, the influence of the model is only locally, but not globally.

To split the estimation region, we first group all observations according to the missing patterns of the essential cancer characters. For example, if there are two essential variables A and B. Then all observations are divided into four groups: 1: both A and B are observed; 2: only A is missing; 3: only B is missing; and 4: both A and B are unknown. Group 1 is defined as an "upper" group of group 2 if the observed essential factors in group 2 are a subset of the observed essential factors in group 1. As in the previous example, group 1 is an "upper" group of groups 2, 3, and 4, and groups 2 and 3 are both "upper" groups of group 4.

As a next step, the categorical and regression tree (CART) is adapted within each group for further split where the binary indicator of missing stage is the response variable and all known essential factors within the group and other related risk factors are predictors. After the split, we expect similar stage distribution within each sub region. Then within each sub region, adaptively use an efficient method to estimate the stage distribution. Specifically, ad hoc method is adapted if the proportion of unknown is very small, say <5%. Weighting method is used if the proportion of unknown is not too much (say, between 5% and 10%). Stabilized weights [21] can be used to avoid very large weights. For large proportion of unknown stages, multiple imputations

are chosen. We allow borrowing information from observations in the "upper" groups that have similar missing pattern when necessary. For example, if the proportion of unknown is large in a leaf, we use the CART to predict all observations in the "upper" groups. The observations that are predicted to be in the same leaf are combined for the estimation of the stage distribution in the leaf.

Finally, the distributions in each leaf are combined to estimate the stage distribution of the whole population. In detail, the following algorithm describes the method

Algorithm 4.1: Adaptive robust method to estimate population distribution:

(1) Split observations to groups according to essential variables' missing status.

(2) For each group:

(a) If the proportion of unknown is very small (<.5%), use ad hoc method to estimate the group distribution, otherwise go to 2b;

(b) Build a tree, f_t, in the group where the binary indicator of missing stage is the outcome and all other variables are predictors;

(c) In each leaf of f_t, follow the procedure described by Figure 1:

(d) Combine the leaf estimates weighted by numbers of observations in leaves to estimate the stage distribution in the group.

(3) Combine the group estimates weighted by numbers of observations in groups to estimate the population stage distribution.

We use CART to split the covariate space into subspaces, in which different methods are chosen to estimate the stage distribution according to the proportion of unknowns.

CART is a binary recursive partitioning algorithm that provides an alternative to traditional parametric models for regression and classification problems. The term "binary" implies that at each step, CART splits a multidimensional covariate space into two regions, and models the response as a constant for each region. Then an optimal variable and split-point are chosen to achieve the best fit again on one or both of these regions. Thus, each node can be split into two child nodes, in which the original node is called a parent node. The term "recursive" refers to the fact that the binary partitioning process can be applied over and over again. Thus, each parent node can give rise to two

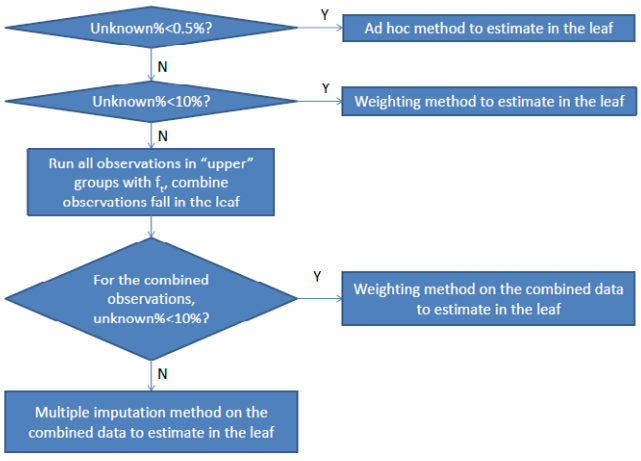

Figure 1: Procedure plot.

child nodes and, in turn, each of these child nodes may themselves be split to generate additional children. CART represents information in a way that is intuitive and easy to be visualized [20].

Build weight/predictive model in algorithm 4.1

The predictive or weight models in the procedure can be chosen according to the needs of data analysis. To estimate the cancer stage distribution for population-based cancer registries, several features of the data present substantial challenges. First, the linear relationship is not adequate to describe the association of the stage distribution with other covariates. Second, complicated interactions might exist among covariates. Third, lots of observations miss one or more values in some covariates. As discussed above, when the sample size is large, it is more beneficial to build nonparametric predictive/weighting models. In this paper, we propose to use multivariate additive regression trees (MART) established by Friedman [22]. MART is an ensemble technique that aims to improve the performance of a single model by fitting many models and combining them for prediction. MART employs two algorithms: "regression tree" from classification and regression tree [23] (CART) and "boosting" that builds and combines a collection of models, i.e. trees [24].

Boosting is one of the recent enhancements to tree-based methods that have achieved considerable success in prediction accuracy. In boosting, models such as regression trees are fitted iteratively to the training data, using appropriate methods to gradually increase emphasis on observations modeled poorly by the existing collection of trees.

Empirical results indicate that MART achieves high accurate prediction performance compared with its competitors. Moreover, compared with the classical parametric regression methods, MART has the following advantages: (1) MART is able to capitalize on the nonlinear relationships between the dependent and independent variables with no need for specifying the basic functions. Unlike many automated learning procedures, which lack interpretability and operate as a "black box", MART provides results that represent valuable tools for interpreting the nature and magnitude of the covariate association with the outcome [25,26]. (2) Due to the hierarchical splitting scheme in regression trees, MART is able to capture complex and/or high order interaction effects. (3) As a tree-based method, MART can handle mixed-type predictors (i.e. quantitative and qualitative covariates) and missing values in covariates. Therefore MART addresses all the challenges in estimating the stage distribution.

In Algorithm 2.1, we use MART to build weighting models, where the binary indicator of missing stage is the response variable and all known essential variables and other factors are explanatory variables. We also use MART in MI to build predictive models based on observations with known stage, where the stage is the outcome and all other variables are covariates in the model.

Estimating the Lung Cancer Stages

We applied the proposed AD method to estimate the lung cancer stage distribution based on 38 population-based cancer registries over the year 2004 to 2009. The data were from the CINA Deluxe Analytic file, collected by North American Association of Central Cancer Registries (NAACCR). NAACCR receives resident cancer case information from its member registries across the US and Canada. For this study, 38 member registries in US signed the active consent form that permitted NAACCR to combine the incidence data from these registries into a single comprehensive data file. The study was based

on the combined cancer incidence data. For more information about CINA Deluxe Analytic file, the reader is referred to the NAACCR website (http://naaccr.org/Research/CINADeluxe.aspx).

The collaborative staging (CS) of lung cancer is extracted from three essential elements: the extension of tumor (CS extension, T), lymph nodes involvement (CS lymph nodes, N), and metastasis condition at diagnose (CS mets at DX, M). If all the essential elements are missing, CS stage cannot be obtained. If any but not all essential variables are missing, CS stage can be known or unknown depending on the values of observed essential factors. There are a total of 997,683 lung cancer cases from the 38 cancer registries over the five years, death certificate only and autopsy only cases excluded. Figure 2 shows the stage distributions by year with and without missing data. Table 1 shows that when different essential variable(s) is unknown, the proportion of unknown stage cases varies. Therefore, we need different analysis strategies when different combinations of essential variables are missing in estimating stage distributions. Meanwhile, we have to use as much information as possible in estimation.

Other factors that potentially relate to missing stage can be grouped to five categories: data collection information such as year of diagnosis, information sources, and confirmation method; demographic information such as age at diagnosis, sex, and race; tumor characters such as tumor size, grade, and histology; treatment types such as surgery, radiation, chemotherapy, and hormone therapy; and census-tract level social-economic information such as employment, education and poverty [1,27]. As examples, Figure 3 shows that the stage distributed differently at different levels of age groups, tumor grades and radiation therapy respectively. To accurately estimate the stage distribution of lung cancer in the population, we should take into account these important factors.

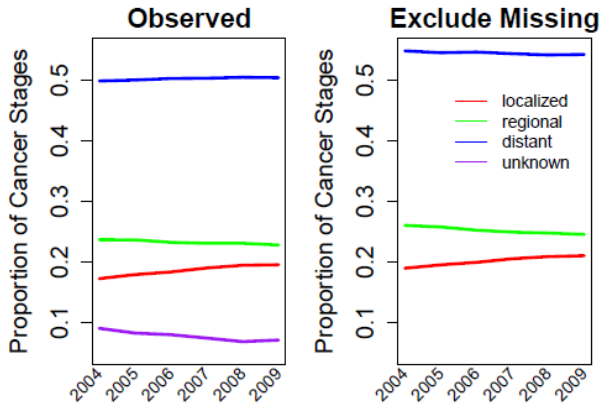

Figure 2: Stage distributions with and without missing information.

CS Lymph Nodes	CS Extension	CS Mets at DX	% Missing in Stage
known	known	known	0
known	known	missing	0.15
known	missing	known	15.44
known	missing	missing	26.49
missing	known	known	1.09
missing	known	missing	5.58
missing	missing	known	13.63
missing	missing	missing	99.97

Table 1: The proportion of unknown stages when different essential variables are missing.

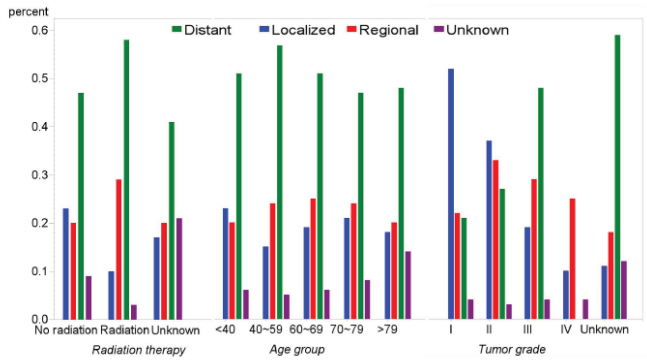

Figure 3: Stage distributions by age groups, tumor grades, and radiation therapy respectively.

Figure 4 compares the estimates of stage distributions by different methods. The imputation methods were implemented using the R package "mice" 8, where predictive mean matching was used for numeric data, logistic regression for binary data, polytomous regression for unordered categorical data and proportional odds model for ordered categorical data with more than two levels. In comparison, we also used the MI with random forest as the predictive model (rf imputation). Five complete data sets were created for analysis. Figure 4 shows that estimates based on the proposed AR method are closer to those from ad-hoc method and from imputation, compared with the weighting and rf imputation.

Also the weighting methods and the r_f imputation show very large variations over the years. By majority voting from different methods, the proportions of localized stages increased, while the proportions of regional and distance staged lung cancer cases decreased over the years. Without knowing the true distributions, we were unable to determine which method is better. To make comparisons, we did a simulation study in 'Simulations section'.

Simulations

This simulation was based on the lung cancer data from the CINA Deluxe Analytic file as described above. The missing data sets were created based on the complete data set where all essential variables were observed and the SEER stages abstracted. We followed the following steps to create the missing data set:

1. Create missing predictors other than the essential variables: The missing in predictors other than the essential variables was created independently for each variable. Each predictor was made to miss at the same rate as in the original data set.

2. Create missing in essential variables: Essential variables were made to miss with probabilities depending on non-essential predictors. There were three essential variables, so there were eight potential patterns of data missing: no missing, missing any one, missing any two, or missing all of the essential variables. Based on the original data set, we built models to calculate the probabilities of an observation having each of the eight missing patterns, where all nonessential predictors were used in building the model. Missing in a predictor was counted as a special category of the predictor. Eight models were built for the missing patterns. The models predicted the probabilities of every missing pattern for each observation in the missing data set created from the first step. A missing pattern was then randomly picked for the observation based on the standardized probabilities.

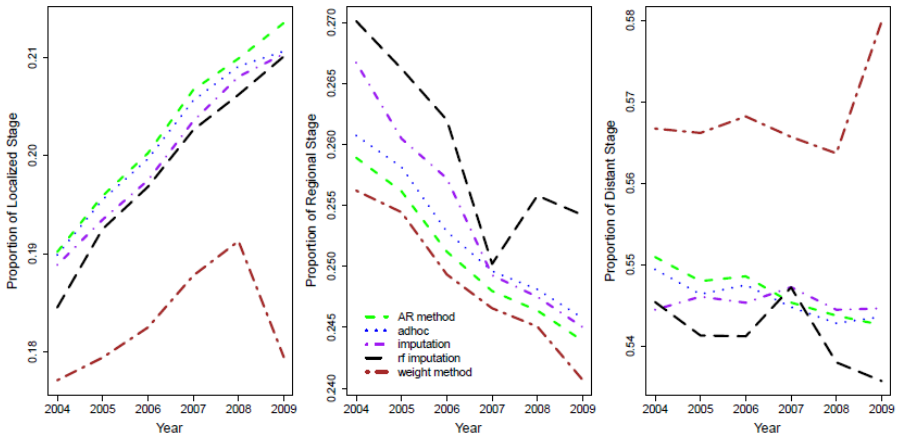

Figure 4: Estimated distributions of lung cancer stages over the year 2004-2009.

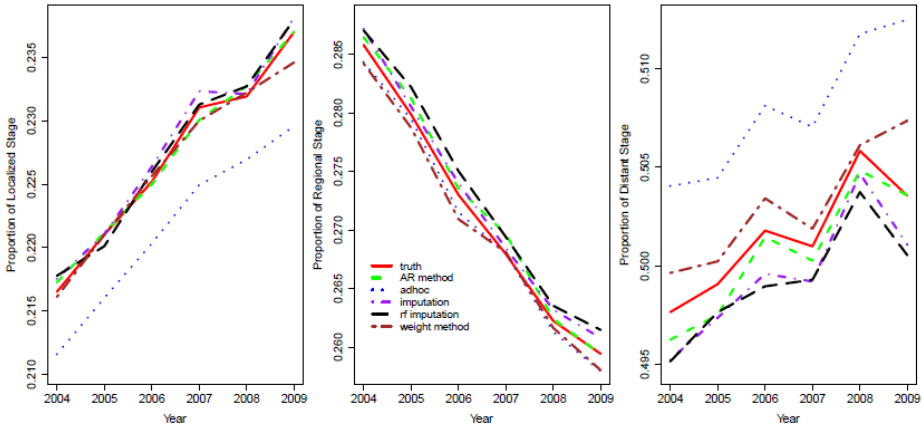

Figure 5: Estimated distributions of lung cancer stages over the years 2004-2009 from simulated data.

3. Create missing in stage: Finally the missing of stage was created based on the data set created from step 2. A model was built based on the original data set where the response variable was the binary indicator of missing stage, and all essential variables and other variables were predictors. Again, missing a predictor was counted as a special category of the predictor. MART was used for the model building. Then the model was used on the data set created from step 2 to generate missing in stage.

The missing data sets were created for each year of 2004–2009 independently. The estimated distribution from the proposed AR methods were compared with the true value from the complete data set, and compared with the estimates from traditional methods such as ad hoc method, weighting method and the multiple imputation method. For the MI method, we used both multinomial logit regression and random forest as the predictive models. We repeated the MI process 50 times and the reported stage distribution was the average of the 50 repetitions. The comparisons of the estimate are shown in Figure 5. The relative sum of absolute errors compared with the AR method, defined as the sum of absolute errors from the corresponding methods divided by that from AR, are shown in Table 2. We found that on average, the proposed method produced much better estimates of the stage distributions over the years, followed by weighting method and multiple imputations. Within multiple imputations, random forest performed a little worse than logistic regression as predictive models. This indicated that linear relationship might be sufficient in describing the associations between stages and other variables. However, the multiple imputation methods were most time consuming as shown in Table 3. Weighting method took least time but was not as stable compared with AR method. All the analyses were carried out in R version 2.8.

Year	Adhoc	MI(logistic)	MI(rf)	Weighting
2004	4.57	1.73	1.79	1.42
2005	3.44	0.77	1.49	0.74
2006	11.86	4.10	5.28	3.90
2007	3.59	1.04	0.99	0.62
2008	5.99	1.16	2.10	0.60
2009	302.35	84.08	102.03	128.90

Table 2: Relative sum of absolute errors (RAE) for each method compared with the AR method. RAE is defined as the AE of the corresponding method divided by that of the AR method. The RAE for the year 2009 is very high since the sum of absolute error for AR is very small (0.006%).

	2004	2005	2006	2007	2008	2009
AR method	810	587	829	849	705	651
Weighting	357	371	387	394	401	371
MI (logistic)	8242	6623	7148	7424	7641	6886
MI (rf)	11406	10655	11153	11287	11879	10899

Table 3: Time in seconds took to get the estimates of proportions of stages in lung cancer.

Conclusion and Future Work

Accurately estimate stage distribution is important in cancer research. Effectively deal with the unknown stage in population-based cancer register center is essential for the estimation. To handle the missing values in population cancer stage is special in that the sample size is big, and the stages are not directly observed but are abstracted from essential tumor information. Depending on the availability of the essential tumor factors, the missed stages can have very different distributions. In the paper, we propose a robust estimation method that adaptively chooses an efficient method to estimate the missing information according to the missing pattern of the essential factors. We apply the AR method to estimate the lung cancer stage distribution over the year 2004 to 2009 for 38 combined major US tumor registries. Simulations show that the proposed method is efficient in estimating the cancer stage in that it consistently gives more accurate estimations while takes acceptable short amount of computing time, when compared with traditional methods such as weighting and multiple imputation. In the AR method, we use MART to build the predictive model for multiple imputations or to estimate weights for weighting method. MART is effective since it can automatically identify significant nonlinear relationship and important interactions. MART is a nonparametric method and it is efficient when the sample size is large, which is especially useful for estimating cancer stage distribution in population. The proposed method can be extended to deal with complex missing data in population distribution estimation.

There are some limitations with the proposed method. We adapt different strategies to handle missing values by stratify the data set according to the proportion of missing. The choice of threshold is ad hoc. A cross-validation method might help to choose the tuning parameters. Also we provide only point estimations in the paper. To make inferences on the point estimates, we can use bootstrap to measure the uncertainties.

All the analyses in the paper were performed with R codes. As a future research, we will write the process with SAS macros. So that when original data set is imputed, and the response variable and essential factors are positioned, the SAS macro will output the estimated distribution of the response variable. In the meantime, the macro will output data sets with weights on known observations and imputations for some observations with missing response. The output data sets with weights will then represent the population.

References

1. Fan Y, Yu Q, Wu X, Hsieh MC (2013) Data Quality Evaluation Using MART Guided Generalized Linear Mixed Model - With Application to Evaluate the SEER Cancer Staging Data. Journal of Data Analysis and Operations Research.
2. Little RA, Rubin, DB (2002) Statistical Analysis with Missing Data (2ndedn.) Wiley-Interscience, New York.
3. Schafer JL, Graham JW (2002) Missing data: our view of the state of the art. Psychol Methods 7: 147-177.
4. Rubin DB (1987) Multiple Imputation for Nonresponse in Surveys. John Wiley and Sons, New York.
5. Horton NJ, Laird NM, Murphy JM, Monson RR, Sobol AM, et al. (2001) Multiple informants: mortality associated with psychiatric disorders in the Stirling County Study. Am J Epidemiol 154: 649-656.
6. Schafer JL (1997) Analysis of Incomplete Multivariate Data. Chapman & Hall, London.
7. Van Buuren S, Groothuis-Oudshoorn K (2011) Mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 45: 1-67.
8. Horton NJ, Kleinman KP (2007) Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. Am Stat 61: 79-90.
9. van Buuren S (2007) Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 16: 219-242.
10. Robins JM, Rotnitzky A, Zhao LP (1995) Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. Journal of the American Statistical Association 90: 106-121.
11. Graham JW, Taylor BJ, Gumsille PE (2001) Planned missing-data designs in analysis of change. New methods for the analysis of change: 335-353.
12. Eisemann N, Waldmann A, Katalinic A (2011) Imputation of missing values of tumour stage in population-based cancer registration. BMC Med Res Methodol 11: 129.
13. Carpenter JR, Kenward MG, Vansteelandt S (2006) A Comparison of Multiple Imputation and Doubly Robust Estimation for Analyses with Missing Data. Journal of the Royal Statistical Society Series A 169: 571-584.
14. Robins JM, Rotnitzky A (2001) Comment on Inference for semiparametric models: some questions and an answer. Statistica Sinica 11: 920-936.
15. Van der Laan MJ, Robins JM (2003) Unified Methods for Censored Longitudinal Data and Causality. Springer, NewYork.
16. Bang H, Robins JM (2005) Doubly robust estimation in missing data and causal inference models. Biometrics 61: 962-973.
17. Kang JDY, Schafer JL (2007) Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. Statistical Science 22: 523-539.
18. Saunders JA, Morrow-Howell N, Spitznagel E, Dor P, Proctor EK (2006) Imputing missing data: A comparison of methods for social work researchers. Social Work Research 30: 19-31.
19. Ibrahim JG, Chen MH, Lipsitz SR, Herring AH (2005) Missing Data Methods for Generalized Linear Models: A Comparative Review. Journal of the American Statistical Association 100: 332-346.
20. Gordon L (2013) Using Classification and Regression Trees (CART) in SAS Enterprise Miner for Applications in Public Health. Data Mining and Text Analytics.
21. Cole SR, Hernán MA (2004) Adjusted survival curves with inverse probability weights. Comput Methods Programs Biomed 75: 45-49.
22. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. The Annals of Statistics 29: 1189-1232.
23. Breiman L, Friedman JH, Olshen R, Stone C (1984) Classification and Regression Trees. Wadsworth, Pacific Grove.
24. Breiman L (1998) Arcing classifiers (with discussion). Ann Statist 26: 801-849.
25. Friedman JH, Meulman JJ (2003) Multiple additive regression trees with application in epidemiology. Stat Med 22: 1365-1381.
26. Yu Q, Li B, Scribner R (2009) Hierarchical Additive Modeling of Nonlinear Association with Spatial Correlations - An Application to related alcohol outlet destruction and changes in neighbourhood rates of assaultive violence. Statistics in Medicine 28: 1896-1912.
27. Hsieh MC, Yu Q, Wu XC, Wohler B, Fan Y, et al. (2012) Evaluating factors associated with unknown SEER Summary Stage 2000 derived from collaborative stage at central registry level. J Registry Manag 39: 101-106.