Journal of Data Analysis & Operations Research
ISSN 2384-4671
J. Data Ana. Oper. Res. 1, No. 1 (Jun. 2014)

Data Quality Evaluation Using MART Guided Generalized Linear Mixed Model with Application to Evaluate the Unknown Stage

Ying Fan[1], Qingzhao Yu[2], Xiao-Cheng Wu MD MPH CTR[3], Mei-Chin Hsieh MSPH CTR[3]

[1]School of Public Health, Louisiana State University Health Sciences Center, New Orleans, LA
[2]Covance at Otsuka Pharmaceutical Development & Commercialization, Inc.
[3]Louisiana Tumor Registry and Epidemiology Program, School of Public Health, Louisiana State University Health Sciences Center, New Orleans, LA

Received: 12 Jun. 2013, Revised: 6 Aug. 2013; Accepted: 14 Dec. 2013, Published: Jun. 2014

Abstract: The stage of cancer is an essential factor for prognosis assessment and treatment planning. Previous studies have found that the proportion of unknown stage cases differed substantially by cancer registry, and identified a list of factors that contribute to the variation of unknown stage among registries. However, only linear relationships between predictors and proportion of unknown stage cases were examined at registry level, which may ignore the individual level information and more complex associations and interactions. The objective of this study is to identify predictors of unknown stage cancer cases and their interactions using MART guided generalized linear mixed model for selected cancers (i.e, colorectal, cervix, female breast, lung, and prostate). Colorectal cancer was used as an example to illustrate the method in this paper. Findings may help registries implement target actions to improve the quality of stage data.

Invasive colorectal cancer cases diagnosed in 2004-2008 from 32 registries that met the North American Association of Central Cancer Registries' (NAACCR) criteria for high quality of incidence data were analyzed. The outcome variable was stage at diagnosis (known stage versus unknown stage). Explanatory variables included individual level demographic variables, type of reporting source, diagnostic confirmation, year of diagnosis, histology type, tumor grade, and registry; as well as county-level variables: poverty, education, and employment status. We first adopted Multiple Additive Regression Trees (MART) to identify significant predictors of unknown stage and interactions, and then used generalized linear mixed model for statistical inference.

Histology type, tumor grade, reporting source and diagnostic confirmation were important factors in predicting unknown staged colorectal cancer. Type of histology interacted with diagnostic confirmation on the risk of unknown stage. After controlling for important factors, a few registries still had significantly higher proportion of unknown staged colorectal cancer cases. Further study is needed to identify the underlying causes.

Keywords: Cancer Stage, Cancer Registry, MART, Generalized Linear Mixed Model

1. Introduction

Cancer stage reflects the extent of the disease at diagnosis. Stage data collected by population-based cancer registries have been used to examine disparities in cancer screening across different populations and areas, monitor trends in stage, and assess the effectiveness of early detection intervention programs. Unfortunately, not all cancer registry cases are staged. Proportions of unknown stage cases vary substantially by registry and cancer sites1. Using the 2004-2007 incidence data from population-based cancer registries, the Data Assessment Work Group of the NAACCR Data Use and Research Committee found that percentage of unknown stage cases varies from 2.4% to 18.8% for colorectal cancer, 2.4% to 18.7% for lung cancer, 1.0% to 13.7% for female breast cancer, and 0.6% to 18.1% for prostate cancer1.

Proportions of unknown stage cases at registry level could be attributable to numerous factors. One of the major factors might be inappropriate coding, which can be improved through training and quality control activities. The large variations in the proportion of unknown stage in different registries may also relate to differences in demographic characteristics of cancer patients, such as residence area and age. Norredam et al. (2008) found that migrant women with breast cancer had a higher probability of diagnosis at unknown stage. Other factors such as type of reporting source and

E-mail: qyu@lsuhsc.edu.

diagnostic confirmation may also be associated with the proportion of unknown stage cases. Recent increases in case ascertainment from non-hospital facilities, such as freestanding pathology laboratories and physician offices, have improved the completeness of cancer incidence data but may also contribute to higher proportion of unknown stage cases. To better understand the issue and set up appropriate actions to reduce the proportion of unknown stage cases, significant factors have to be determined.

In the phase-one analysis with the Data Assessment Work Group, we examined the associations of unknown staged cases with selected factors for four different cancers using linear regression models. We found that factors associated with unknown stage differed by cancer sites. Type of reporting source was a significant predictor of unknown stage for all cancers except for lung cancer after adjusting for other variables (Hsieh et al. 2012). However, linear models are not sufficient to describe all patterns of the associations and interactions among contributing factors. For example, we found, in a pilot testing, that type of reporting source and diagnostic confirmation were both highly correlated with unknown stage (p-value<0.05), but the effects were not additive; the interaction effect between them was also significant. It is difficult to use the linear regression models to detect all possible interactions. Another weakness of linear regression is that, to deal with missing data problems, an ad hoc method in linear regression is to delete all observations that have one missing variable, which may result in the loss of information and biased results (Greenland and Finkle, 1995). To overcome the limitations, we proposed this phase-two analysis.

To adopt a two-stage analysis, we first used the multiple additive regression tree (MART) (Fritz et al. 2000), a nonlinear model, to identify factors and interactions that were significantly associated with unknown stage. The inference methods with MART were used to estimate the proportion of contributions of the selected factors in explaining the variations in reporting unknown stage cancer cases and to identify significant interactions. As a second stage, the detected significant factors and interactions were included in a generalized linear mixed model for statistical inference. Finally, we evaluated the quality of tumor registry stage data by comparing the residual proportions of unknown stage cases after controlling for significant factors and interactions. In this paper, we used colorectal cancer (ICD-O-3: C18.0-C18.9, C19.9, C20.9) (Friedman, 2001) as an example to demonstrate the proposed method.

2. Data and Methods

2.1 Data Source

The 2004-2008 data from population-based cancer registries that met NAACCR's high quality criteria for incidence data and gave consent to use the data for this study were included. Autopsy and death-certificate-only cases were excluded from the analysis. The binary outcome variable was whether or not an invasive cancer case was diagnosed with unknown SEER Summary Stage 2000 derived from Collaborative Stage Version 1 (CSv1). Explanatory variables included in the analysis were demographic variables of the patients (i.e., race, gender, and age), clinical variables (i.e., diagnostic confirmation, histology, and tumor grade), year of diagnosis, and type of reporting source. Type of histology was grouped as neoplasms, Not Otherwise Specified (NOS), epithelial neoplasms NOS, adenocarcinoma NOS, and specific histologies. Type of reporting source was categorized as hospital and non-hospital. The hospital category included hospital inpatient, radiation treatment centers/medical oncology centers, and hospital outpatient units/surgery centers. The non-hospital category included physician's office, nursing home/hospice, and laboratory only.

For confidentiality purposes, each registry was assigned a unique random number ranging from 1 to 32 so that no registry would be identifiable.

2.2 Statistical Analysis

MART is a data-based strategy that has been advocated for exploring variable relationships. Compared with the classical parametric regression methods, MART has the following advantages: (1) MART is able to capture the nonlinear relationships between the dependent and independent variables with no need for specifying the basis functions. (2) Because of the hierarchical splitting scheme in regression trees, MART is able to capture complex and/or high-order interaction effects. (3) Unlike many automated learning procedures, which lack interpretability and operate as 'black box', MART provides tools to interpret the nature and magnitudes of covariate relations with the outcome (for example, relative variable importance and partial dependence plot in Fritz et al. (2000), as described below). (4) MART can handle mixed-type predictors (i.e. quantitative and qualitative covariates) and missing values in covariates. (5) MART has shown a superior exploration and prediction performance in epidemiology research, see (Friedman and Meulman, 2003) and (Yu and Scribner, 2009).

One of the most important aspects of model interpretation involves the identification of the relative importance of covariates in terms of their relative strength in predicting the response, and understanding their joint effects on the response. For tree-based methods, Breiman et al. (1984) proposed a measure of importance $I_j^2(b_H)$ for each variable x_j, based on the number of times that variable was selected for splitting in the tree b_H weighted by the squared improvement to the model as a result of each of those splits. Friedman (2001) generalized this importance measure to additive tree expansions by taking the average over the trees $I_j^2 = \frac{1}{M}\sum_{m=1}^{M} I_j^2(b_H)$. The measure turns out to be more reliable than a single tree as it is stabilized by averaging. Since these measures are relative, we scale the measure so that the importance of all the variables sum to 100 percent. In addition to the importance measure, Friedman and Meulman (2003) also introduced a concept called partial dependence to describe the dependence of the fitted model on a subset of variables. Given any subset x_s of the input variables indexed by $s \in \{1, \dots, p\}$ the partial dependence is defined as $F_x(x_s) = E_{xs}[f(x)]$ where $E_{xs}[\cdot]$ means expectation over the joint distribution of all the input variables not indexed in s.

In this study, unknown stages (as a binary variable, with logit link) were regressed on potential covariates by MART. Relative importance of each variable and potential interaction was calculated to identify important factors. Partial dependence plots were drawn to show how factors were related to unknown staged cases and were used to guide variable transformation. After important variables and interactions were identified by MART, generalized linear mixed model (GLMM) was adopted for further exploration. Mixed model was used to analyze data with complex patterns of variability (Snijders et al. 1999). The binary response variable in GLMM was whether or not a cancer case was unstaged. The predictors were those variables and interactions identified by MART and were transformed so that they are reasonably linearly associated with the outcome (log odds of being staged as unknown). The level II variable in GLMM was registry. If interactions involving registry were identified as important by MART, the predictors that were interactively effective with registries were treated as covariate with random effect in GLMM.

3. Results

Variable relative importance

We demonstrated the proposed statistical method on the unknown staged cases for colorectal cancer. Table 1 lists the variable relative importance in predicting the probability of unknown stage. Histology type was the most important factor, followed by tumor grade, type of reporting source and diagnostic confirmation. Age, year of diagnosis, race and sex had relative low impact on unknown stage.

Table 1. Variable Relative Importance for Colorectal Cancer

Rank	Risk Factor	Relative Importance (%)
1	Histology	38.10
2	Grade	21.81
3	Report Souce	15.27
4	Confirmation	12.47
5	Registry	8.09
6	Age of Diagnosis	3.95
7	Year of Diagnosis	0.22
8	Race	0.07
9	Sex	0.01

Partial dependence plots

Figure 1 represents the partial dependent plots of predictors. The predictors were recoded as follow:

$$\text{Grade} = \begin{cases} 0 & \text{differentiated;} \\ 1 & \text{moderately differentiated;} \\ 2 & \text{Poorly differentiated;} \\ 3 & \text{undifferentiated;} \\ 4 & \text{unknown grade.} \end{cases}$$

$$\text{Histology} = \begin{cases} 0 & \text{specific histologies} \\ 1 & \text{neoplasms, NOS;} \\ 2 & \text{epithelial neoplasms, NOS;} \\ 3 & \text{adenocarcinoma, NOS} \end{cases}$$

$$\text{Confirm} = \begin{cases} 0 & \text{microscopic;} \\ 1 & \text{non-microscopic;} \\ 2 & \text{unknown.} \end{cases}$$

$$\text{Race} = \begin{cases} 0 & \text{white;} \\ 1 & \text{black;} \\ 2 & \text{others.} \end{cases}$$

$$\text{Source} = \begin{cases} 0 & \text{hospital;} \\ 1 & \text{non-hospital facilities.} \end{cases}$$

$$\text{Sex} = \begin{cases} 0 & \text{male;} \\ 1 & \text{female.} \end{cases}$$

Age ranging from 1 to 85 years old was regrouped with 5 years intervals and treated as continuous variable.

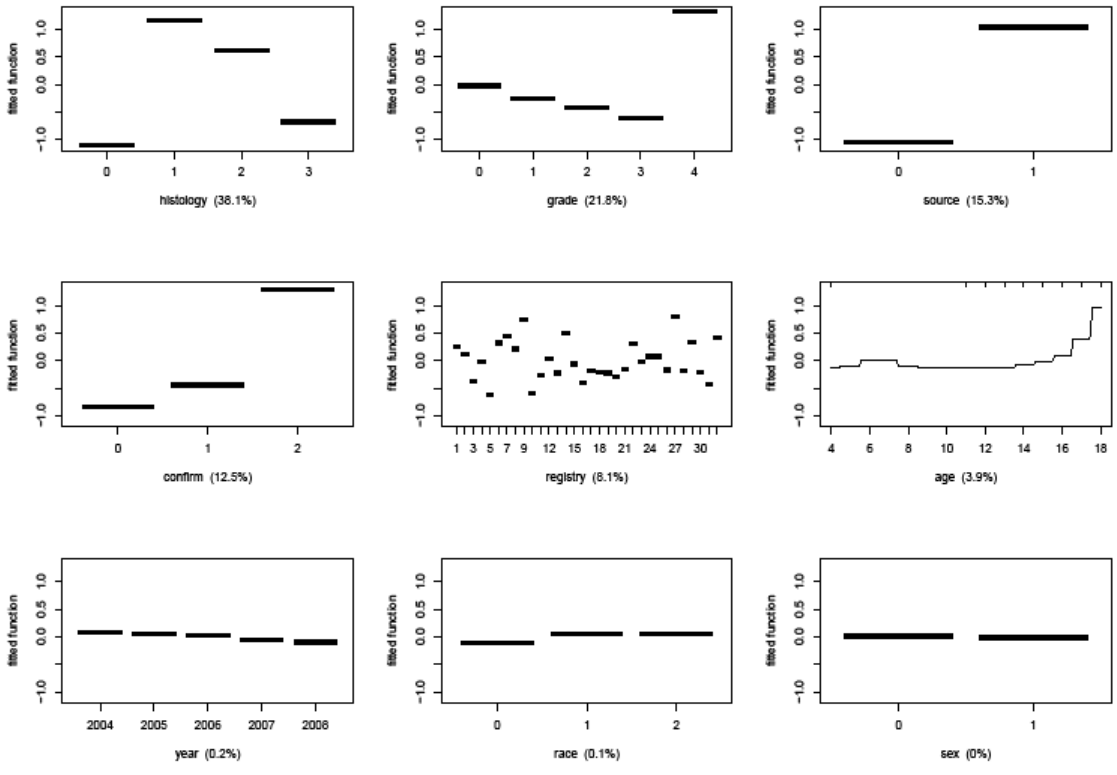

Figure 1 Partial Dependent Plots for Colorectal Cancer

Figure 1 indicates that colorectal cancer cases diagnosed with specific histologies were related to the lowest probability of unknown stage. Those diagnosed as neoplasms NOS, however, were most likely to miss the stage information among all histology types. Microscopic or non-microscopic confirmed colorectal cases had lower probability of unknown stage than the ones without microscopically confirmed. Cases were more likely to have unknown stage if they were reported from non-hospital facilities, compared with cases reported from hospital facilities. Colorectal cancer cases with unknown grade had significantly higher probability of unknown stage compared to cases with other types of grade. It is notable that well differentiated colorectal tumors were more likely to be diagnosed as unknown stage compared to moderately, poorly differentiated or undifferentiated tumor. The partial dependence plot of registry suggests that on average, registries 9 and 27 had relatively higher proportion of unknown staged colorectal cancer cases than other registries. Averaged over all registries, the probability of unknown stage was constant when the diagnosis age was younger than 79, then the probability climbed up after 80. The years of diagnosis, patient's race and gender appeared to have negligible relationship to unknown stage.

Important interaction

The two most important pairwise interaction effects were registry by reporting source, and histology type by diagnostic confirmation. That means the association between reporting source unknown stage varied across registries. Moreover, histology related to unknown stage in various ways for different confirmation types.

Table 2. Variable interaction for Colorectal Cancer

Interaction Term	Size
Registry∗Report source	43.07
Histology∗Confirmation	42.96

Based on MART analysis result, we examined the relationships between the probability of unknown stage and predictors, histology type, grade, reporting source and diagnostic confirmation using generalized linear mixed effect model. More specifically, registry was the level II variable and the effect of reporting source on unknown stage was treated as random effect in the GLMM model. Histology type, tumor grade and diagnostic confirmation were fixed effects, and the interaction between histology type and diagnostic confirmation was also included in the model.

Generalized linear mixed effect model

The generalized linear mixed effect model is presented as follows:

$$Log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_{0j} + \beta_{1j}X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 X_3 + \beta_5 X_4$$

where

$$\beta_{0j} = \gamma_{00} + \mu_{0j}$$
$$\beta_{1j} = \gamma_{10} + \mu_{1j}$$

$\mu_{0j} \sim N(0,\ \sigma_0^2)$ and $\mu_{1j} \sim N(0,\ \sigma_1^2)$.

X_1, X_2, X_3, and X_4 denote reporting source, diagnostic confirmation, histology, and tumor grade with values defined previously. Random intercept μ_{0j} represents the variation in the average responses among registries, and fixed intercept γ_{00} denotes the overall mean probability of unknown staged colorectal cancer in logit scale. Similarly, the random slope μ_{1j} describes the variation in effect of report source on unknown stage, and fixed slope γ_{10} measures the overall effect of report source on unknown stage. Table 3 represents the fixed effect and covariance parameter estimates.

The estimated fixed intercept, $\hat{\gamma}_{00}$, was -3.80 indicating that the overall odds of unknown stage for a colorectal cancer case was $e^{-3.80} = 0.022$, given all other predictors are at reference levels. Overall, the odds of having unknown stage for a colorectal cancer case reported from non-hospital was $e^{1.79} = 5.99$ times the ones from hospital after controlling for the remaining factors. Cases confirmed by non-microscopic examination or without confirmation information had significantly greater probability of being diagnosed as unknown stage than microscopically confirmed cases, with odds ratio 1.38 (CI: [1.22,1.57]) and 7.31 (CI: [6.14,8.71]), respectively. Compared to specific histologies, cases with neoplasm NOS, epithelial and adenocarcinoma histology type were more likely to be unknown staged. The estimated odds ratios were 3.38 (CI: [2.65,4.32]), 2.19 (CI: [1.71,2.80]), and 1.40 (CI: [1.08,1.81]), respectively. Compared to well differentiated, invasive colorectal tumor with unknown grade information was more likely to be diagnosed with unknown stage (OR=3.69, CI:[3.50,3.89]). However, moderate, poorly

differentiated or undifferentiated tumor was less likely to be unknown staged, with estimated odds ratios 0.73 (CI:[0.69,0.77]), 0.63 (CI: [0.59,0.67]), and 0.51 (CI: [0.43,0.61]), respectively.

The estimated variance of random intercept, $\hat{\sigma}_0^2$, tells the variation among registries. It was estimated at 0.08 with standard error 0.04. The test of variance parameter based on the residual pseudo-likelihood indicated that σ_0^2 was significantly greater than 0, which suggested that the probability of unknown staged colorectal cancer varied among registries. The estimated random effect for registry showed that after controlling for reporting source, diagnostic confirmation, histology type and tumor grade, registries 9 and 29 still had significantly higher proportion of unknown staged colorectal cancer case. The significance of $\hat{\sigma}_1^2$ suggested that reporting source related to unknown stage differently among registries. For example, compared to cases reported from hospital facilities, those from non-hospital facilities in registry 27 were less likely to be unknown staged. However, non-hospital reported cases were more likely to be unknown staged in registries 6 or 29.

Table 3. Model estimates for Colorectal Cancer

Effect	Estimate (SE)	P-value
non-hospital	1.79 (0.09)	<0.0001
unknown confirmed	2.64 (0.27)	<0.0001
non-microscopic confirmed	1.52 (0.22)	<0.0001
neoplasms, NOS	2.42 (0.05)	<0.0001
epithelial neoplasms, NOS	1.71 (0.03)	<0.0001
adenocarcinoma, NOS	0.68 (0.02)	<0.0001
unknown confirmed*neoplasms	-1.35 (0.29)	<0.0001
unknown confirmed*epithelial	-1.04 (0.30)	0.0006
unknown confirmed*adenocarcinoma	-0.24 (0.31)	0.44
non-microscopic confirmed*neoplasms	-2.25 (0.23)	<0.0001
non-microscopic confirmed*epithelial	-1.73 (0.23)	<0.0001
non-microscopic confirmed*adebcarcinoma	-0.79 (0.24)	0.001
unknown grade	1.30 (0.03)	<0.0001
poorly differentiated	-0.46 (0.03)	<0.0001
moderately differentiated	-0.31 (0.03)	<0.0001
Undifferentiated	-0.67 (0.09)	<0.0001
Covariate Parameter Estimate		
Parameter	**Subject**	**Estimate (SE)**
Intercept	Registry	0.08 (0.04)
Report source	Registry	0.12 (0.03)

4. Discussion

In this study, we used a data-oriented machine learning model to confirm that histology type, tumor grade, reporting source and diagnostic confirmation were important factors that played important roles in predicting unknown staged colorectal cancers. We also observed that diagnostic confirmation had a significant interaction with histology type. However, reporting source was not an important factor on unknown stage if the data were analyzed by the statistical method described in Hsieh et al. (2000).

This study incorporated the advantages of MART and mixed effect model to identify the important factors and evaluate their effects in predicting unknown staged colorectal cancer. Another advantage of MART is its ability to investigate, complex relationship between predictors and response variable as well as interactions among predictors, avoiding the main drawback of simple linear regression model. Compared with multiple regression model with fixed effects, the inclusion of random effects in GLMM allows us to further examine the variation of the effect of predictor on unknown stage across registries, thus providing more insight on how to improve data quality at different registries. In addition, the findings can be generalized to registries not included in the data.

The computational complexity of MART is the main limitation in this study. For this colorectal cancer data, with over 40 million cases, it took several days to build a MART model.

In summary, this study identified that histology type, diagnosis confirmation, reporting source and tumor grade are important predictors of unknown stage for colorectal cancer. There existed an interaction between histology type and diagnostic confirmation in predicting unknown stage. Reporting source related to unknown stage in different manners for different registries. After controlling for these factors, the probability of unknown stage still varied by registries. A future study examining more potential factors associated with unknown stage should be done to further explore the variation among registries.

Acknowledgement

This research was supported by National Cancer Institute grant HHSN261200900015C.

References

Breiman, L. Friedman, J.H., Olshen, R. A., & Stone, CJ. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics, 29*(5), 1189-1232.

Friedman, J. H., Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine, 22*(9), 1365-1381.

Fritz, A., Percy, C., Jack A, Shanmugaratnam K, Sobin L, & Parkin DM. (2000). *ICD-O-3: International classification of diseases for oncology*, 3rd ed. Geneva: World Health Organization.

Greenland, S., & Finkle, W.D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analysis. *American Journal of Epidemiology, 142*, 1255-1264.

Hsieh MC., Yu Q., Wu X.C., Fan Y., Qiao B., Jemal A., & Ajani UA. (2012). Evaluating Factors Associated with Unknown SEER Summary Stage Derived from Collaborative Stage. *Journal of Registry Management, 39*(3), 101-106.

Norredam, M., Krasnik, A., Pipper, C., & Keiding, N. (2008). Differences in stage of disease between migrant women and native Danish women diagnosed with cancer: results from a population-based cohort study. *European Journal of Cancer Prevention, 17*(3), 185-190.

Snijders, Tom A.B., & Bosker, R. J. (1999). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling*. London etc.: Sage Publishers

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 neighborhood rates of assaultive violence. *Statistics in Medicine, 28*(14), 1896-1912.