# Data Quality Evaluation Using MART Guided Generalized Linear Mixed Model – With Application to Evaluate Cancer Staging Data

Ying Fan[1] MS; Qingzhao Yu[1] PhD; Xiao-Cheng Wu[2] MD, MPH; Meichin Hsieh[2] MPH
1: Biostatistics Program, School of Public Health; 2: Louisiana Tumor Registry
This study is partially supported by NAACCR CINA Research Award # HHSN261200900015C

## Introduction

Accurate information on stage the cancer is essential for evaluating prognosis and planning treatment. The NAACCR Data Use and Research Committee's Data Assessment Workshop has found that the percentage of unknown stage varied substantially by cancer registry, and numbers of factors may contribute to the variation[1].

However, the workgroup only examined linear relationships between predictors and unknown stage at registry level, which were insufficient to capture nonlinear patterns of associations and interactions among predictors. To accurately describe all types of associations, statistical methods for applying to both linear and nonlinear relationships need to be explored.

## Objective

This study examines predictors of unknown stage and their interactions using MART guided generalized linear mixed model.

## Method

Data were from the NAACCR CINA Analytic data file including 32 cancer registries. We included invasive colorectal cancer cases diagnosis in 2004-2008. Death certificate only and autopsy only cases were excluded.

$$\text{Histology} = \begin{cases} 0 & \text{specific} \\ 1 & \text{neoplasms, NOS;} \\ 2 & \text{epithelial neoplasm;} \\ 3 & \text{adenocarcinoma, NOS} \end{cases} \qquad \text{Grade} = \begin{cases} 0 & \text{differentiated;} \\ 1 & \text{moderately differentiated;} \\ 2 & \text{poorly differentiated:} \\ 3 & \text{undifferentiated;} \\ 4 & \text{unknown grade} \end{cases}$$

$$\text{Confirm} = \begin{cases} 0 & \text{microscopic:} \\ 1 & \text{non-microscopic;} \\ 2 & \text{unknown} \end{cases} \qquad \text{Race} = \begin{cases} 0 & \text{white:} \\ 1 & \text{black:} \\ 2 & \text{others} \end{cases}$$

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

The binary response variable is whether the cancer case was staged as unknown (y=1) or otherwise (y=0) at time of diagnosis.

Multiple Additive Regression Trees (MART) [2]method is adapted to identify important factors and interactions. MART produces partial dependence plots, which are used to guide the transformation of important factors for reasonable linear associations with the unknown stage.

The transformed predictors and interactions are used in generalized linear mixed models for further inferences. The predictors that have important interaction with registry enter the mixed model as random effects.

## Results

**Figure 1: Relative Importance of Predictors**

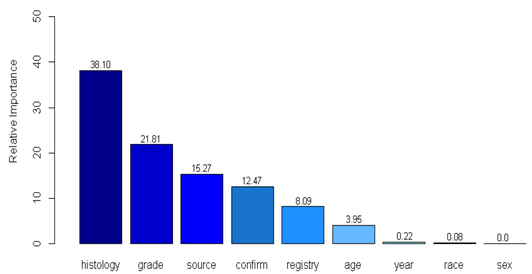
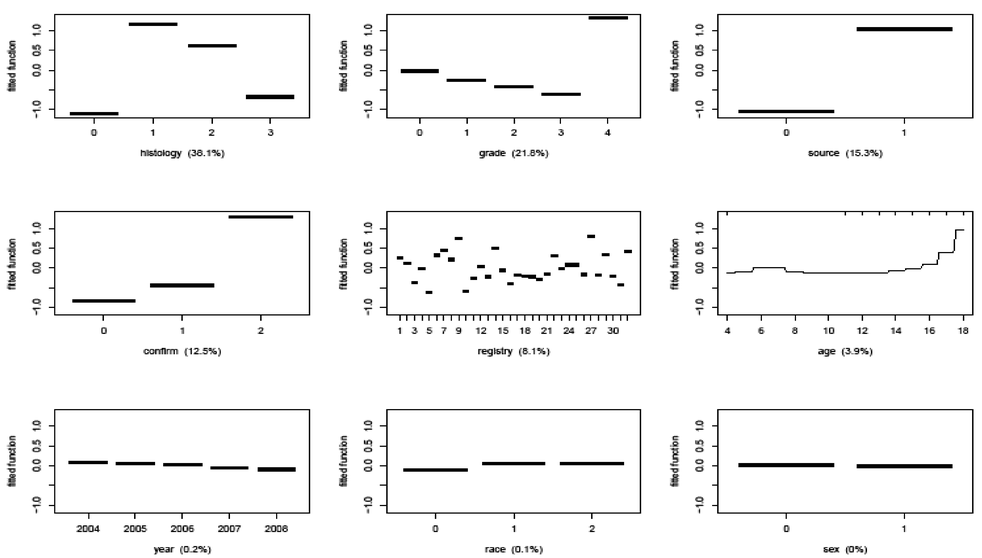



**Table 1: Important Interaction Terms**

| Interaction Term | Size |
|---|---|
| **Registry * Report Source** | 43.07 |
| **Histology * Confirm** | 42.96 |

The most important interactions are registry by report source, and histology by confirmation

**Figure 2: Partial Dependence Plots**




## Generalized Linear Mixed Model

$$\text{Log}\left(\frac{\pi_{ij}}{1-\pi_{ii}}\right) = \beta_{0j} + \beta_{1j} \cdot \text{Report} + \beta_2 \cdot \text{Confirm} + \beta_3 \cdot \text{Histology} + \beta_4 \cdot \text{Confirm} * \text{Histology} + \beta_5 \cdot \text{Grade}$$

$$\beta_{0i} = \gamma_{00} + \mu_{0i}, \qquad \mu_{0i} \sim N(0, \sigma_0^2)$$

$$\beta_{1i} = \gamma_{10} + \mu_{1i}, \qquad \mu_{1i} \sim N(0, \sigma_1^2)$$

**Table 2: Random Effect**

| Covariance Parameter Estimate | | |
|---|---|---|
| **Parameter** | **Subject** | **Estimate (SE)** |
| **Intercept** | Registry | 0.08 (0.04) |
| **Report Source** | Registry | 0.12 (0.03) |

**Figure 3: Caterpillar Plot for Predicted Logit Response with 95% CI**

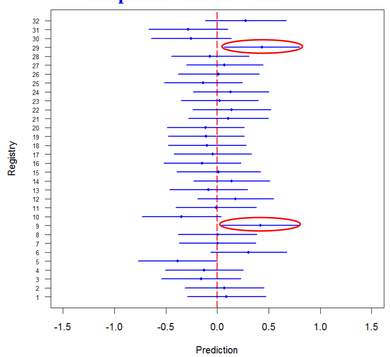



**Table 3: Fixed Effect Contrasts**

| Contrast | OR (95% CI) |
|---|---|
| **Histology 1 vs Histology 0** | 3.38 (2.65,4.32) |
| **Histology 2 vs. Histology 0** | 2.19 (1.71,2.80) |
| **Histology 3 vs. Histology 0** | 0.40 (1.08,1.81) |
| **Grade 1 vs. Grade 0** | 0.73 (0.70,0.77) |
| **Grade 2 vs. Grade 0** | 0.63 (0.59,0.67) |
| **Grade 3 vs. Grade 0** | 0.51 (0.43,0.61) |
| **Grade 4 vs. Grade 0** | 3.69 (3.50,3.89) |
| **Confirm 1 vs. Confirm 0** | 1.38 (1.22,1.57) |
| **Confirm 2 vs. Confirm 0** | 7.31 (6.14,8.71) |

Conclusion:

1. Histology type, tumor grade, report source and diagnosis confirmation were important factors in predicting unknown staged colorectal cancer.
2. Registry 9, 29 had significant higher proportions of unknown staged colorectal cancer cases after controlling for important factors.
3. The association between report source and unknown stage varied for different registries.

Reference:
1. Hsieh, MC., Yu, Q., Wu, X.C., Wohler, B., Fan, Y., Jamison, M., Umed, A. (2012). "Evaluating Factors Associated with Unknown SEER Summary Stage 2000 Derived from Collaborative Stage." submitted to Journal of registry management.
2. Friedman, J.H. (2001). "Greedy function approximation: a gradient boosting machine". The Annals of Statistics. Vol. 29, No. 5, 1189-1232}