

Statistical Models for Hospital Readmission Prediction with Application to Chronic Obstructive Pulmonary Disease (COPD) Patients

Li Zeng, Smriti Neogi and Jamie Rogers

**Department of Industrial and Manufacturing Systems Engineering
The University of Texas at Arlington, Arlington, TX, 76019, USA**

Susan Seidensticker, Carlos Clark, Lindsey Sonstein, Rick Trevino and Gulshan Sharma

**Department of Internal Medicine
The University of Texas Medical Branch, Galveston, Texas, 77555, USA**

Abstract

Hospital readmission is an important health care performance indicator for quality monitoring. Many efforts have been made to identify risk factors to readmission and build prediction models in various medical applications. However, there is a lack of rigorous studies on the issues in the statistical analysis for readmission modeling and guidelines on how to choose appropriate models fitting specific cases. To bridge the gap, this paper reviews five popular statistical models that can be used to model binary readmission outcomes, and provides general strategies/guidelines for model construction. These methods are illustrated in a case study using a dataset from chronic obstructive pulmonary disease (COPD) patients.

Keywords

Hospital readmission, Statistical modeling, Logistic regression, Bayesian estimation, Chronic obstructive pulmonary disease (COPD)

1. Introduction

With concerns regarding the rising cost of health care, hospital readmission shortly after discharge, which is common and costly, has attracted considerable attention in recent years. Readmission is perceived as an adverse effect in itself, and suggested to be used as health service performance indicator for quality monitoring (Kansagara et al. 2011; Ross et al. 2008). One typical example is the readmission of chronic obstructive pulmonary disease (COPD) patients. COPD is the third leading cause of death in the United States and the only leading cause for which morbidity and mortality are rising (Kuo et al. 2009). The natural course of the disease is marked by acute exacerbations requiring frequent interactions with the whole spectrum of the health care system — from outpatient to emergency room to inpatient care. Over half of COPD patients who are hospitalized for acute exacerbations are readmitted at least once in the subsequent 6 months. Hospitalization for such patients accounts for as much as 40% of the total direct cost of medical care for COPD in the nation (Cao et al. 2006).

Many studies have been done to identify risk factors and build prediction models for hospital readmission in various medical applications (e.g., Chen and Narsavage 2006; Kariv et al. 2006; Stewart 2000; Zitser-Gurevich 1999). In those studies, readmission is typically measured either by binary indicators of whether a patient had readmission within a short period after discharge, e.g., 30 days, or by the interval between discharge and readmission. This paper focuses on the modeling of binary readmission measurements. The most popular statistical method used in such studies is the logistic regression (LR) model, which is a special type of generalized linear models commonly used for binary outcomes (Myers et al. 2002). Only a few studies consider special issues in model construction, such as variable selection (Cao et al. 2006; Ferraris et al. 2001) and correlation in the outcomes (Hasan et al. 2010). In general, however, there is a lack of rigorous studies on the issues in the statistical analysis for readmission modeling and guidelines on how to choose appropriate models fitting each specific case. Moreover, advanced statistical and data mining methods for binary outcomes that have been widely used in other fields should be considered in readmission studies to improve the performance of current models.

This study aims to bridge the gap by providing a review of available statistical models which can be used for readmission modeling in different situations and general strategies/guidelines for model selection, variable selection and parameter estimation in model building. These methods are illustrated in a case study using data from an ongoing study of the authors on COPD patients.

The available statistical models for binary outcomes can be divided into two categories depending on the characteristics of the data: models for *independent* outcomes, and models for *correlated* outcomes. When there is only one observation from each patient, the binary outcomes are typically assumed to be independent; when there are multiple observations from the same patients due to repeated admissions of those patients, the outcomes of the same patients will be treated as correlated data. It needs to be pointed out that when it is believed that the correlation between the outcomes from the same patients is moderate or the number of patients with multiple admissions is relatively small, the independence assumption is more appropriate to avoid the complexity in characterizing the correlation structure of the outcomes. In this paper, we will review five popular statistical models in the two cases: regular LR model, Bayesian LR (BLR) model, and Logistic regression tree (LRT) model for independent outcomes, and generalized estimating equations (GEE) and generalized linear mixed model (GLMM) for correlated outcomes.

The remainder of this paper is organized as follows. Models and model construction strategies for independent outcomes and correlated outcomes will be given in Section 2 and 3 respectively. Section 4 will present the results of the case study, and Section 5 will conclude this paper and discuss potential issues in using these methods.

2. Statistical Models for Independent Outcomes

Let $\{(y_i, \mathbf{X}_i), i=1, \dots, n\}$ be the observation of patient i , where y_i indicates whether the patient was readmitted within the period of interest ($=1$) or not ($=0$), $\mathbf{X}_i$ is the vector of available covariates including patient risk factors (e.g., age, comorbidities) and health care process variables (e.g., treatments, interventions), and n is the total number of observations. Note that $\mathbf{X}_i$ may also contain functions of individual covariates (e.g., quadratic terms, logarithm), or interactions of covariates, and thus often has a high dimension. The basic task in readmission prediction is to build a statistical model to characterize the dependency of readmission on the covariates and identify statistically significant covariates. In the follows, three statistical models that can serve this purpose will first be briefly introduced, and then the proposed strategy for model construction will be given.

Logistic regression model

In the LR model, the binary readmission outcomes are assumed to follow a Binomial distribution

$$y_i \sim \text{Binomial}(p_i) \quad (1)$$

where p_i is the probability of readmission of patient i . This probability depends on the covariates through

$$\log \frac{p_i}{1-p_i} = \boldsymbol{\beta} \mathbf{X}_i \quad (2)$$

where $\boldsymbol{\beta}=[\beta_1, \dots, \beta_p]'$ are the parameters of this model which represent the effects of covariates. These parameters are usually estimated using maximum likelihood estimation methods. The LR model is used extensively in modeling binary outcomes in many disciplines, especially medical and social science fields.

Bayesian logistic regression model

The Bayesian LR model takes the same model form as the LR model, i.e., (1)~(2), except that the parameters are viewed as random variables instead of fixed, unknown quantities ,

$$\begin{aligned} \boldsymbol{\beta} &\sim \pi(\boldsymbol{\beta}) \\ \boldsymbol{\beta} | D &\sim P(\boldsymbol{\beta} | D) \propto \pi(\boldsymbol{\beta}) \times L(D) \end{aligned} \quad (3)$$

where D denotes data, $\pi(\boldsymbol{\beta})$ is called the prior distribution which represents prior information/expert knowledge of the parameters, $L(D)$ is the likelihood which represents the information contained in the data, and $P(\boldsymbol{\beta} | D)$ is the posterior distribution of the parameters which combines the prior information on the parameters and information in the data. The prior distribution can be elicited from expert opinions or estimated from historical data. When no prior information is available, uniform priors will be used, that is, equal probability is assumed for every point in the parameter space. The posterior distribution typically has no analytical form and needs to be simulated using Markov chain Monte Carlo (MCMC) techniques. With the simulated sample from the posterior distribution, point estimates of the parameters can be obtained using summaries, e.g., mean, median or mode, of the sample.

Bayesian models have been widely used in the modeling of discrete outcomes like the binary readmission outcomes (Gelman, et al. 2004). Spiegelhalter (2004) suggests to use such models in healthcare evaluation to complement conventional statistical methods. Besides the ability to incorporate expert knowledge, the BLR model has three other attractive features for readmission modeling: (1) Unlike conventional statistical methods, Bayesian approaches does not rely on asymptotics and can apply for any sample size. For this reason, Bayesian models are preferred when sample size is small, which is often the case in readmission studies (the typical sample size in the literature is several hundred). (2) The output of Bayesian estimation is the posterior distribution of the parameters which provides much more information on the parameters than the point estimates from conventional estimation methods. (3) From the simulated sample of the posterior distribution, credible intervals of each parameter can be obtained, which is the analogy of the confidence intervals in conventional estimation. The credible intervals bear a more intuitive interpretation than the confidence intervals as the range of the parameters under certain confidence level.

Logistic regression tree model

The LRT model developed by Chan and Loh (2004) is a tree-structured extension and generalization of the LR model, which is a counterpart of the linear regression tree model for continuous data. The basic idea of such models is to partition the sample space into subspaces and build LR models with simple structures within each subspace,

$$\log \frac{p_i}{1-p_i} = \begin{cases} \beta_I X_i & \text{if } X_i \in \text{subspace I} \\ \beta_{II} X_i & \text{if } X_i \in \text{subspace II} \\ \dots & \dots \end{cases} \quad (4)$$

Usually categorical covariates are used for the partition and continuous covariates are used in fitting the LR models. For example, patients may be divided into groups by their gender and/or race, and one LR model is fitted for each group. The main advantage of LRT models is that it can overcome the interpretability issues of the LR model in the face of multicollinearity, nonlinearity and interactions, without sacrificing estimation accuracy. Another advantage lies in the intuitive graphical representation of the model structure, as illustrated in Figure 5 in the case study. The LRT model can be constructed through the LOTUS (Logistic Tree with Unbiased Selection) algorithm by Chan and Loh. Please see Chan and Loh (2004) for more information about this tool.

Model construction strategy

There are two main issues in building the readmission model:

- *Selection of appropriate models.* The above three models have their pros and cons: the LR model is simple, but lacks of easy interpretation in complex cases; the BLR model works better for small sample sizes, but the specification of prior distributions will be difficult and computation time will increase in the presence of high-dimensional covariates; the LRT model bears better interpretability, but this good property will be affected when the model has many covariates and the tree structure is very complex. As is the case in any regression analysis, there is no “best” model for a given dataset, and it is useful to consider all possible ways of explaining the data and choose one that fits the goal of the study. A general strategy is: if the model involves only a small number of covariates, the LRT model is preferable; otherwise, the LR or BLR model is better. In the latter case, to integrate the strength of the LR and BLR model, we can first build a LR model to establish the model structure, and then use Bayesian approaches to estimate model parameters.
- *Selection of variables.* There is typically a large pool of covariates available in the readmission modeling, including the main effects, interactions and other functions of individual risk factors and process variables. Variable selection needs to be conducted to identify significant covariates as well as build a model with good prediction performance.

Considering the above issues, a general strategy for model construction is suggested as follows:

Step 1: Variable screening: Fit simple LR models for each individual covariate, that is, the considered covariate is the only covariate in the LR model, and retain significant covariates for the next step. This is equivalent to a bivariate analysis for each covariate. The purpose of this screening is to identify all the potential significant covariates from the pool of available covariates.

Step 2: LR modeling: Build a LR model using all the significant covariates and conduct model selection through stepwise procedures based on likelihood ratio test or information criteria such as the Akaike information criterion (AIC). The resulting model will be used in a finer analysis in the following step.

Step 3: BLR and LRT modeling: Build a BLR model and a LRT model based on the model from last step. Compare the two models and choose the one that better fits the specific study.

3. Statistical Models for Correlated Outcomes

Let $\mathbf{y}_i = [y_{i1}, \dots, y_{it}]'$ be the outcomes of patient i , where t is the number of observations from this patient and can vary from patient to patient. These outcomes are assumed to be correlated as they are from the same patient. Such correlation needs to be taken into account in building the readmission model. Statistical models for repeated and longitudinal measurements can be used for this purpose. Two popular ones of them are GEE (Liang and Zeger 1986) and GLMM (van Montfort et al. 2010), which have been widely used in biomedical and epidemiological studies (e.g., Sashegyi et al. 2000; Ghahroodi et al. 2010; Williamson et al. 1996; Lipsitz and Fitzmaurice 1996).

Generalized estimating equations

The GEE method takes the same model form as the LR model, with an incorporation of the correlation structure of outcomes from the same patients, which leads to an estimator of the parameters β given by (Myers et al. 2002)

$$\mathbf{V} = A_i^{1/2} \mathbf{R} A_i^{1/2} \quad (5)$$

$$\sum_{i=1}^n \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \mathbf{V}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

where $\boldsymbol{\mu}_i$ is the mean vector which is a function of $\boldsymbol{\beta}$, $\mathbf{R}$ is a working correlation matrix of $\mathbf{y}_i$ which is common to all patients, $\mathbf{V}$ is the corresponding covariance matrix, and A_i is a $t \times t$ diagonal matrix with the variance of $\boldsymbol{\mu}_i$ as the diagonal elements. The working correlation matrix $\mathbf{R}$ needs to be specified for the estimation. Popular choices include: (1) *Independent*: All correlations are assumed to be zero. (2) *Exchangeable*: All correlations are assumed to be equal. (3) *1-dependent*: Every observation is only correlated with adjacent observations. (4) *Unspecified*: All correlations are estimated from data. The GEE estimates can be obtained via the Newton-Raphson algorithm. One good property of this method is that it yields consistent estimates even when the correlation structure is misspecified.

Generalized linear mixed model

The basic idea of GLMM is to characterize the heterogeneity across patients by assuming the regression coefficients to be random and follow a certain probability distribution. Under this assumption, the outcomes of the same patients are correlated as they share the same unobserved coefficient. For this reason, the GLMM is often used to handle correlations in the outcomes. Specifically, the simplest form of the logistic mixed model is

$$\log \frac{p_i}{1-p_i} = \boldsymbol{\beta} \mathbf{X}_i + u_i \quad (6)$$

where $\boldsymbol{\beta}$ is the fixed effect of covariates, and u_i is the random effect of patient i , $i=1, \dots, n$, which are assumed to be independent and identically distributed as $N(0, \sigma^2)$. σ^2 is the variance of the random effect, which is a measure of the patient heterogeneity. In this model, the correlation of outcomes from the same patient, i.e., $y_{i1}, \dots, y_{it}$, is a constant depending on σ^2 . The model in (6) is more precisely referred to as a random-intercept model as the random effect is on the intercept. The random effect can also occur on the coefficients of some predictors, e.g., $\beta_k + \delta$, where β_k is the fixed effect of X_k , and δ is the associated random effect. The estimators of $\boldsymbol{\beta}$ and σ^2 can be obtained by the restricted maximum likelihood estimation method.

Model construction guidelines

Whether to choose GEE or GLMM in modeling correlated outcomes depends on the specific situation and goal of the study. General guidelines are:

- When the concern is the population-averaged readmission, GEE should be used, while when the concern is the heterogeneity among patients in readmission, GLMM should be used (Hu, et al. 1998).
- In GLMM, the correlation of outcomes from the same patients is a constant for all patients, while in GEE, there are other options. So GEE is able to characterize more complex correlation structures. On the other hand, however, more data will be required in GEE to estimate the desired correlation matrix.

4. Case Study

The models described in Section 2 and 3 have been applied to data from an ongoing study of the authors on COPD patients. The dataset contains information of 420 discharges during 2009~2012. The outcome of interest is the readmission of patients within 30 days after discharge, and data of 36 potential covariates are collected, covering patient characteristics (e.g., demographics, comorbidities, and habitual behaviors), care process variables (e.g., steroid usage, treatment, procedural intervention usage), and previous healthcare utilization (e.g., the number of emergency room visits and hospitalizations). The observations are from 282 patients, among whom 220 (78%) have

only one observation, 41 (14.5%) have two observations, and 21 (7.5%) have more than two observations. Using these data, two studies have been done: in *Study 1*, methods in Section 2 are applied, assuming the observations are independent, and in *Study 2*, methods in Section 3 are applied considering the correlation among observations from the same patients. All the models are constructed using R, a free statistical software, except the LRT model which is built using the LOTUS algorithm. Results of the two studies are given in Section 4.1 and 4.2 respectively.

4.1 Results in Study 1

Following the suggested model construction procedure in Section 2, a preliminary analysis is first done to screen the available covariates by fitting LR model for each of them individually. Significant covariates identified (p -value=0.05) are listed in the first row of Table 1, where the first 5 variables are indicators of whether the patient is a alcohol user (Alcohol), drug user (Drug), had ordered long-acting antimuscarinic agents (LAMA), had flu vaccine (Flu), and discharged with oxygen (Oxygen), and the other 3 variables are the number of previous emergency room visits (ER), hospitalizations (HOS) and hospitalizations for COPD (HOSC). The p -values indicate that ER and HOS have most significant effects. Then such analysis is applied to two-way interactions of these significant variables. Significant interactions are listed in the second row in Table 1. The identified main effects and interactions will be used in building the three types of models for independent outcomes.

Table 1: Significant covariates identified in the preliminary analysis

main effects	Alcohol(0.045) Drug(0.0191) LAMA(0.0142) Flu(0.0231) Oxygen(0.000284) ER(3.14e-10) HOS(8.53e-07) HOSC(0.00027)
2-way interactions	Alcohol*Drug(0.012) Alcohol*Flu(0.0019) Alcohol*Oxygen(0.0002) Alcohol*ER(2.5e-06) Alcohol*HOS(1.7e-05) Alcohol*HOSC(6e-04) Drug*Flu(0.0011) Drug*Oxygen(0.002) Drug*ER(0.0006) Drug*HOS(0.0013) Drug*HOSC(0.004) LAMA*Oxygen(0.001) LAMA*ER(0.002) LAMA*HOS(0.0022) LAMA*HOSC(0.003) Flu*Oxygen(0.0009) Flu*ER(1.62e-07) Flu*HOS(4.96e-06) Flu*HOSC(0.0003) Oxygen*ER(2.9e-08) Oxygen*HOS(5.3e-07) Oxygen*HOSC(0.0002)

Fitted LR Model

The LR model is built through a stepwise procedure based on likelihood ratio tests,

$$\begin{aligned} \log[p/(1-p)] = & -7.75 + 1.18 \times \text{Alcohol} + 0.91 \times \text{Drug} + 0.70 \times \text{LAMA} + 2.89 \times \text{Flu} + 5.20 \times \text{Oxygen} \\ & + 3.09 \times \text{ER} + 0.08 \times \text{HOS} + 0.76 \times \text{HOSC} - 1.35 \times \text{Alcohol} \times \text{Oxygen} - 0.98 \times \text{Drug} \times \text{ER} \\ & - 2.59 \times \text{Flu} \times \text{Oxygen} - 0.81 \times \text{Flu} \times \text{ER} - 1.15 \times \text{Oxygen} \times \text{ER} - 0.80 \times \text{Oxygen} \times \text{HOSC} \end{aligned}$$

Significant covariates in the model are indicated by bold. One characteristic of this model is that all the main effects are positive, while all the interactions are negative, which makes it difficult to interpret the effect of each variable. As one example, Figure 1 shows the main effect of Flu and the interaction effect of Flu and Oxygen. From the main effect plot, we can see that having flu vaccine will increase the probability of readmission. According to the interaction plot, however, this is only the case for patients discharged without oxygen; for those with oxygen, flu vaccine seems to cause no considerable difference. LR models with both main effects and interaction effects may produce complex results like this which are hard to interpret or counterintuitive.

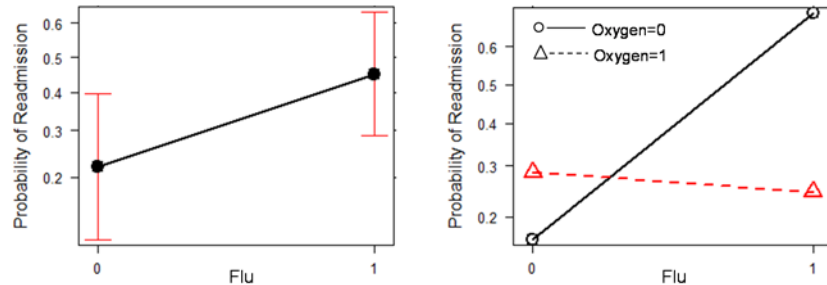


Figure 1: Effect of Flu (left) and Flu*Oxygen (right)

Fitted BLR Model

Building a BLR model here means estimating parameters in the LR model via Bayesian approaches. This is done by using the *MCMClogit* function in the R package *MCMCpack*, which generates one sample from the posterior

distribution of parameters based on data and specified priors. For simplicity, here we use the default noninformative priors for all the parameters, which corresponds to no prior information/expert knowledge. To be robust, the median of the posterior sample of each parameter is taken as point estimate of the parameter. The fitted model is

$$\begin{aligned} \text{logit}[p/(1-p)] = & -8.63 + 1.34 \times \text{Alcohol} + 0.86 \times \text{Drug} + 0.68 \times \text{LAMA} + 3.40 \times \text{Flu} + 5.96 \times \text{Oxygen} \\ & + 3.46 \times \text{ER} + 0.08 \times \text{HOS} + 0.84 \times \text{HOSC} - 1.53 \times \text{Alcohol} * \text{Oxygen} - 0.99 \times \text{Drug} * \text{ER} \\ & - 3.02 \times \text{Flu} * \text{Oxygen} - 0.95 \times \text{Flu} * \text{ER} - 1.29 \times \text{Oxygen} * \text{ER} - 0.88 \times \text{Oxygen} * \text{HOSC} \end{aligned}$$

The Bayesian estimates are higher than the LR estimates for some covariates, and lower for others. Figure 2 displays the estimates of the 15 covariates in the LR model and the BLR model, which shows their difference more clearly.

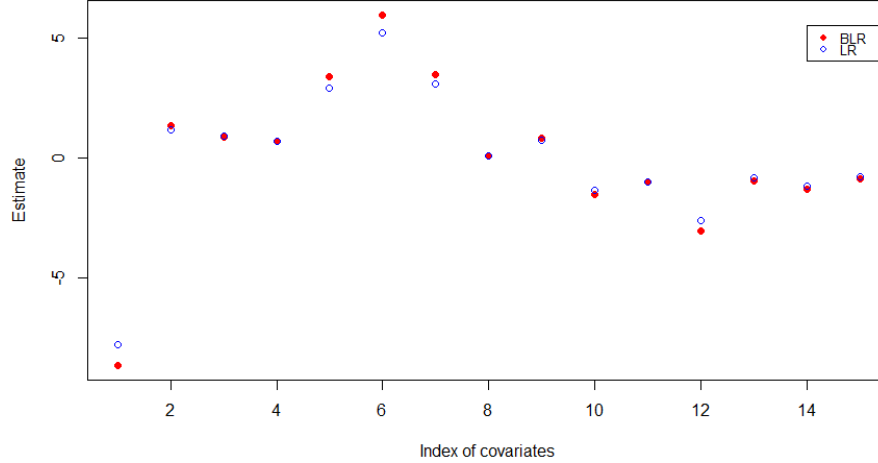


Figure 2: Parameter estimates in the LR model and BLR model

From the sample of the posterior, the empirical distribution of the posterior can be obtained. As an example, the lower panel of Figure 3 shows the posterior distribution of the parameter associated with Flu, where the left plot is the trace of the posterior sample, and the right plot is the density function of the posterior estimated using kernel methods. Clearly, the posterior does not follow normal distribution. To show the effect of sample size on the posterior distribution, we have also fitted a BLR model using only 150 observations. The resulting posterior sample and density function are shown in the upper panel of Figure 3. We can see that this distribution has larger variance, i.e., with a flatter shape, than the posterior distribution based on all the 420 observations. This is consistent with a defining feature of Bayesian estimation: as sample size becomes larger, the estimation will become more accurate.

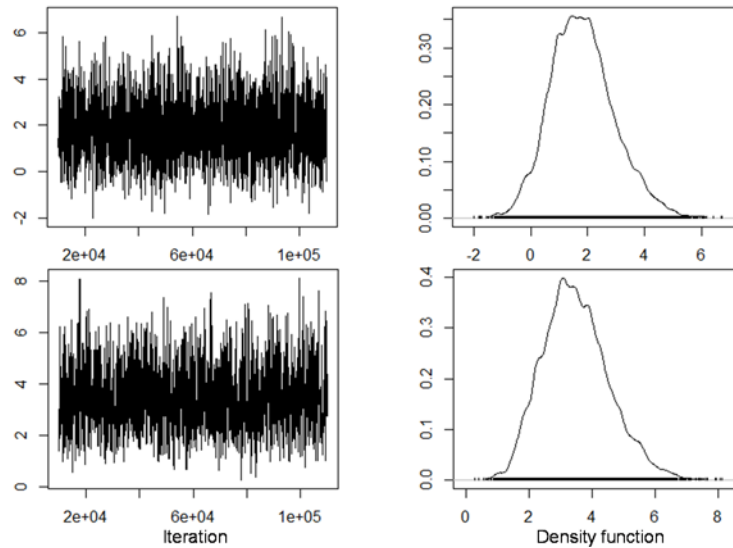


Figure 3: Posterior sample and distribution: $n=150$ (upper) and $n=420$ (lower)

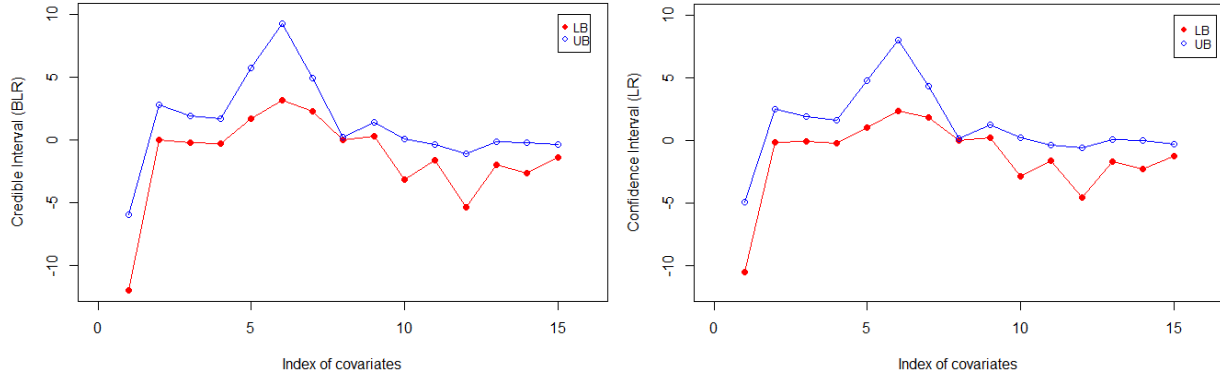


Figure 4: Credible intervals of BLR model (left) and confidence intervals of LR model (right)

The credible interval of the Bayesian estimates can be obtained simply by finding the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of the posterior sample, where α is the significance level specified to be 0.05 in this study. For comparison, we have also calculated the corresponding confidence intervals of the LR model. The lower bound (LB) and upper bound (UB) of the intervals for each covariate are shown in Figure 4. We can see that the BLR intervals and the LR intervals have similar shape, but bear little difference for some covariates. Moreover, due to the small sample size, some intervals are relatively large, indicating low precision.

Fitted LRT Model

A LRT model containing the selected variables (i.e., Alcohol, Drug, LAMA, Flu, Oxygen, ER, HOS, HOSC) is constructed through a stepwise procedure (p -value = 0.05). The 5 categorical variables are used to split the tree and the other 3 continuous variables are used to fit a LR model in each branch. The graphical representation of the resulting LRT model is shown in Figure 5. In the figure, the circles and squares represent intermediate and terminal nodes, respectively. The number inside a node is the node label and the splitting rule of an intermediate node is given beside it. If a case satisfies the rule, it goes to the left child node; otherwise the right child node. The ratio of cases with $y=1$ to the node sample size is given beneath each terminal node. For example, the ratio “21/209” below node 2 means that there are 209 cases with “Oxygen=0” and 21 of them have readmissions. A LR model is fitted for each terminal node. The fitted model at each terminal node is also given in Figure 5.

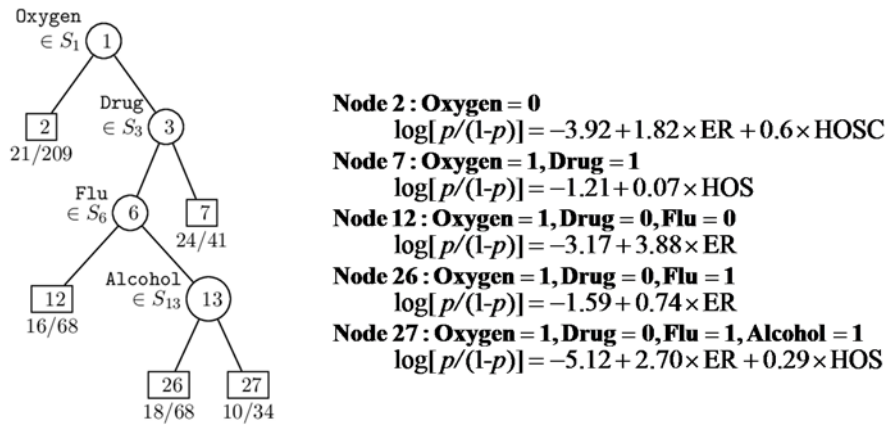


Figure 5: Fitted LR tree (left) and resulting models (right)

The tree model has intuitive interpretation on the dependency of readmission by splitting the whole patient population into subpopulations and providing a simple model to characterize the pattern of each subpopulation. For example, the LRT in Figure 5 suggests that in most cases readmission only or mainly depends on ER. The ratios

below each terminal node also indicate that patients discharged with oxygen and have regular drug use (Node 7) have highest probability of readmission.

4.2 Results in Study 2

The first step in fitting the GEE model is to specify the correlation structure. As there is only a small percentage of patients who have more than one observations, a simple correlation structure will be appropriate. So we choose the “exchangeable” option, which assumes that the correlations between observations of the same patients are equal. For convenience, we still follow the model structure as used in the LR model. The coefficient estimates, standard errors and associated p -values are listed in Table 2, together with their counterparts in the LR model for comparison.

Table 2: Estimates of the LR, GEE and GLMM models

	<u>Estimate</u>			<u>Std. Error</u>			<u>p-value</u>		
	LR	GEE	GLMM	LR	GEE	GLMM	LR	GEE	GLMM
Intercept	-7.75	-7.74	-7.88	1.42	2.10	1.52	5.19e-08	0.0002	2.3e-07
Alcohol	1.18	1.18	1.13	0.67	0.81	0.72	0.0790	0.1489	0.1160
Drug	0.91	0.91	0.81	0.52	0.43	0.61	0.0788	0.0351	0.1843
LAMA	0.70	0.71	0.64	0.48	0.41	0.55	0.1440	0.0813	0.2400
Flu	2.89	2.90	2.83	0.96	1.44	1.02	0.0027	0.0434	0.0055
Oxygen	5.20	5.19	5.03	1.44	2.07	1.56	0.0003	0.0121	0.0012
ER	3.09	3.09	3.33	0.65	1.01	0.76	1.81e-06	0.0023	1.3e-05
HOS	0.08	0.08	0.09	0.05	0.05	0.05	0.1005	0.1354	0.0864
HOSC	0.76	0.76	0.78	0.26	0.26	0.28	0.0038	0.0041	0.0052
Alcohol*Oxygen	-1.35	-1.33	-1.41	0.79	0.89	0.86	0.0884	0.1344	0.1019
Drug*ER	-0.98	-0.96	-0.92	0.31	0.44	0.65	0.0021	0.0279	0.1575
Flu*Oxygen	-2.59	-2.59	-2.71	1.00	1.42	1.07	0.0096	0.0678	0.0113
Flu*ER	-0.81	-0.81	-0.70	0.46	0.57	0.55	0.0785	0.1574	0.2041
Oxygen*ER	-1.15	-1.16	-0.93	0.59	0.80	0.72	0.0509	0.1447	0.1968
Oxygen*HOSC	-0.80	-0.79	-0.78	0.26	0.25	0.27	0.0020	0.0017	0.0044

The results in Table 2 show that the parameter estimates of the GEE model are similar to those from the LR model. However, there is considerable difference in the estimated standard errors. In most cases the standard error from the GEE model is larger than those from the LR model. That is because GEE takes correlation between outcomes into account. Correspondingly, the p -values from these two models are also different. As a result, some significant covariates in the LR model are less significant in the GEE model (e.g., Alcohol and Flu). In other words, the GEE model is able to identify the variables that truly have significant effects on the readmission. The estimated correlation coefficient in the GEE model is -0.0128 with standard error=0.95, which shows low precision. This is likely due to the small number of patients with multiple observations in the dataset.

In building the GLMM, individual random effect in each parameter are considered. The estimated variance of the random effect is close to zero for all the covariates except the main effect of EV. The parameter estimates of the GLMM with random effect in the parameter of EV are listed in Table 2. We can see that the estimates of the GLMM show larger difference from the LR estimates than the GEE model. The standard errors and p -values of the GLMM are between those of the LR model and the GEE model in most cases. The estimated variance of the random effect is 0.974, which means that the main effect of EV varies from patient to patient with a variance of 0.974 (mean of the effect is 3.09 as given in the table).

5. Conclusions and Discussion

This study reviews available statistical models for readmission prediction. For independent readmission outcomes, the logistic regression model, Bayesian logistic regression model, and logistic regression tree model are suggested, while for correlated outcomes, the generalized estimating equations and generalized linear mixed model are suggested. Guidelines on model selection and how to choose between these methods are also provided. The use of these methods is demonstrated in a case study using data from COPD patients. It is found that COPD readmission is related with patient behaviors (alcohol use and drug use), treatment (use of long-acting antimuscarinic agents, flu vaccine), severity (oxygen use at discharge), and health care utilization. Especially, the number of previous

emergency room visits appears to be the most significant factor to readmission. The findings are consistent with some existing studies on this topic (e.g., Smith et al. 2000; Almagro et al. 2006; Bahadori and FitzGerald, 2007).

Two final notes on the use of the models: First, in analyzing a given dataset in practice, we suggest to first check the correlation of observations from the same patients to determine whether to model the outcomes as independent or correlated. This can be realized by fitting a GEE model and checking the correlation estimate. If the correlation estimate is very small, we can simply treat the data as independent data. Second, due to the high dimension of the covariates in readmission studies, accurate model estimation can only be obtained when adequate observations are available. This is especially the case for the GEE model and GLMM. In the case study, the models have been built based on a limited dataset from an ongoing project. As more data becomes available, we will refit these models using the updated dataset and report our findings.

Acknowledgement

The authors gratefully acknowledge financial support from the University of Texas System Chancellor's Health Fellow for Systems Engineering Grant.

References

- Kuo, Y.-F., Sharma, G., Freeman, J., and Goodwin, J. S., Growth of general internal medicine hospitalists caring for older patients in US hospitals, *N Eng J Med*, vol. 360, no. 11, pp. 1102-12, 2009.
- Cao, Z., Ong, K. C., Eng, P., Tan, W. C., and Ng, T. P., Frequent hospital readmissions for acute exacerbation of COPD and their associated factors, *Respirology*, vol. 11, pp. 188-195, 2006.
- Smith, D. M., Giobbie-Hurder, A., Weinberger, M., Oddone, E. Z., Henderson, W. G., Asch, D. A., Ashton, C. M., Feussner, J. R., Ginier, P., Huey, J. M., Hynes, D. M., Loo, L., and Mengel, C. E., Predicting non-elective hospital readmissions: a multi-site study, *Journal of Clinical Epidemiology*, vol. 53, pp. 1113-18, 2000.
- Almagro, P., Barreiro, B., de Echagüen, A. O., Quintana, S., Carballeira, M. R., Heredia, J. L., and Garau, J., Risk factors for hospital readmission in patients with chronic obstructive pulmonary disease, *Respiration*, vol. 73, pp. 311-17, 2006.
- Bahadori, K., and FitzGerald, J. M., Risk factors of hospitalization and readmission of patients with COPD exacerbation – systematic review, *International Journal of COPD*, vol. 2, no. 3, pp. 241-51, 2007.
- Hasan, O., Meltzer, D. O., Shaykevich, S. A., Bell, C. M., Kaboli, P. J., Auerbach, A. D., Wetterneck, T. B., Arora, V. M., Zhang, J., and Schnipper, J. L., Hospital readmission in general medicine patients: a prediction model, *J Gen Intern Med*, vol. 25, no. 3, pp. 211-19, 2010.
- Ferraris, V. A., Ferraris, S. P., Harmon, R. C., and Evans, B. D., Risk factors for early hospital readmission after cardiac operations, *The Journal of Thoracic and Cardiovascular Surgery*, vol. 122, pp. 278-86, 2001.
- Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., and Kripalani, S., Risk prediction models for hospital readmission—a systematic review, *JAMA*, vol. 306, no. 15, pp. 1688-98, 2011.
- Ross, J. S., Mulvey, G. K., Stauffer, B., Patlolla, V., Bernheim, S. M., Keenan, P. S., and Krumholz, H. M., Statistical models and patient predictors of readmission for heart failure, *Arch Intern Med*, vol. 168, no. 13, pp. 1371-86, 2008.
- Chen, Y.-J., and Narsavage, G. L., Factors related to chronic obstructive pulmonary disease readmission in Taiwan, *Western Journal of Nursing Research*, vol. 28, no. 1, pp. 105-24, 2006.
- Kariv, Y., Wang, W., Senagore, A. J., Hammel, J. P., Fazio, V. W., and Delaney, C. P., Multivariate analysis of factors associated with hospital readmission after intestinal surgery, *The American Journal of Surgery*, vol. 191, pp. 364-71, 2006.
- Stewart, R. D., Campos, C. T., Jennings, B., Lollis, S. S., Levitsky, S., and Lahey, S. J., Predictions of 30-day hospital readmission after coronary artery bypass, *Ann Thorac Surg*, vol. 70, pp. 169-74, 2000.
- Zitser-Gurevich, Y., Simchen, E., Galai, N., and Braun, D., Prediction of readmissions after CABG using detailed follow-up data, *Medical Care*, vol. 37, no. 7, pp. 625-36, 1999.
- Myers, R. H., Montgomery, D. C., and Vining, G. G., *Generalized Linear Models—with Applications in Engineering and the Sciences*, John Wiley&Sons, New York, 2002.
- Liang, K.-Y., and Zeger, S. L., Longitudinal data analysis using generalized linear models, *Biometrika*, vol. 73, no. 1, pp. 13-22, 1986.
- van Montfort, K., Oud, J., and Satorra, A. (eds.), *Longitudinal Research with Latent Variables*, Springer, New York, 2010.

- Sashegyi, A. I., Brown, K. S., and Farrell, P. J., Application of a generalized random effects regression model for cluster-correlated longitudinal data to a school-based smoking prevention trial, *American Journal of Epidemiology*, vol. 152, no. 12, pp. 1192-1200, 2000.
- Ghahroodi, Z. R., Ganjali, M., and Kazemi, I., Models for longitudinal analysis of binary response data for identifying the effects of different treatments on insomnia, *Applied Mathematical Sciences*, vol. 4, no. 62, pp. 3067-82, 2010.
- Williamson, D. S., Bangdiwala, S. I., Marshall, S. W., and Waller, A. E., Repeated measures analysis of binary outcomes: applications to injury research, *Accid. Anal. and Prev.*, vol. 28, no. 5, pp. 571-79, 1996.
- Lipsitz, S. R., and Fitzmaurice, G. M., Estimating equations for measures of association between repeated binary responses, *Biometrics*, vol. 52, pp. 903-12, 1996.
- Chan, K.-Y., and Loh, W.-Y., LOTUS: An algorithm for building accurate and comprehensible logistic regression trees, *Journal of Computational and Graphical Statistics*, vol. 13, pp. 826-52, 2004.
- Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R., and Pentz, M. A., Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes, *American Journal of Epidemiology*, vol. 147, no. 7, pp. 694-703, 1998.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B., *Bayesian Data Analysis*, 2nd Edition, Chapman & Hall/CRC, Boca Raton, 2004.
- Spiegelhalter, D. J., Incorporating Bayesian ideas into health-care evaluation, *Statistical Science*, vol. 19, no. 1, pp. 156-74, 2004.

Biography

Li Zeng, PhD is an Assistant Professor in the Department of Industrial and Manufacturing Systems Engineering at the University of Texas at Arlington. Dr. Zeng's research interests are mainly on statistical modeling, monitoring and analysis of complex production and service processes/systems. The goal of her research is to characterize baseline system performance, detect changes in performance and identify root causes for changes. She has applied her methodologies to various complex systems including manufacturing systems and healthcare delivery systems such as cardiac surgeries.

Smriti Neogi is a PhD candidate in the Department of Industrial and Manufacturing Systems Engineering at the University of Texas at Arlington. She is a Lean Six Sigma Black Belt and holds CPIM certification from APICS. She also holds a Healthcare Management Engineering certificate from the Institute of Industrial Engineers (IIE).

Jamie Rogers, PhD, PE is a Professor and Associate Chair of the Department of Industrial and Manufacturing Systems Engineering at the University of Texas at Arlington. Dr. Roger's research interests include the design and analysis of manufacturing and service systems, logistics, sustainability, strategic planning, engineering education, and infrastructure/integration issues relating to agile virtual enterprises. She currently serves on the UT System - Systems Engineering in Healthcare Steering Committee (2011-present) whose mission is to apply industrial and systems engineering tools and methods to further advance clinical effectiveness and safety by improving operations across the health institutions of The University of Texas System.

Susan Seidensticker, MSHAI has 15+ years of experience providing healthcare performance and quality improvement assistance in clinical and nonclinical areas of hospitals nationwide. Her Industrial Engineering background in workflow and process flow analysis, data analysis and team facilitation will be utilized in this research.

Carlos Clark, DO is an Assistant Professor in the Department of Internal Medicine and recently named CMIO. Dr Clark has worked with Epic functionalities since its implementation at UTMB. He has successfully developed and implemented disease specific best practice alerts and smartsets. He currently leads the Epic Workflow Committee to improve the clinical application of Epic functionality.

Lindsay Sonstein, MD is an Assistant Professor and Director of Quality Improvement in the Department of Internal Medicine. She has completed Clinical Safety and Effectiveness training at The University of Texas MD Anderson Cancer Center. She oversees and develops QI projects that align with Institutional priorities throughout the

Department of Medicine. She has extensive experience in developing and implementing Epic smartsets into clinical workflow.

Rick Trevino has 15+ years' experience in healthcare information technology at UTMB. Utilizing skills in web application development, database design (SQL) and systems analysis, Rick has created custom applications that have reduced costs and increased efficiencies. As an expert of the EMR (EPIC) data repository, his technical knowledge enables him to produce meaningful data summaries.

Gulshan Sharma, MD, MPH is an Associate Professor and Director Division of Pulmonary and Critical Care Medicine. He is currently funded by NIA-NIH to investigate continuity of care across transition using Medicare data. His research examines the role of provider continuity across transitions from the community to the hospital and back again and its impact on quality of medical care. He is also a recipient for Health IT grant from UT system on improving care coordination through EHR. He is a nationally recognized expert on studying the importance of continuity of care.