A DECISION SUPPORT TOOL FOR PREDICTING PATIENTS AT RISK OF READMISSION: A COMPARISON OF CLASSIFICATION TREES, LOGISTIC REGRESSION, GENERALIZED ADDITIVE MODELS, AND MULTIVARIATE ADAPTIVE REGRESSION SPLINES

Eren Demir

University of Hertfordshire Business School Working Paper (2012)

The Working Paper Series is intended for rapid dissemination of research results, work-in-progress, and innovative teaching methods, at the pre-publication stage. Comments are welcomed and should be addressed to the individual author(s). It should be noted that papers in this series are often provisional and comments and/or citations should take account of this.

University of Hertfordshire Business School Working Papers are available for download from https://uhra.herts.ac.uk/dspace/handle/2299/5549 and also from the British Library: www.mbsportal.bl.uk

Copyright and all rights therein are retained by the authors. All persons copying this information are expected to adhere to the terms and conditions invoked by each author's copyright. These works may not be re-posted without the explicit permission of the copyright holders.

The Business School at the University of Hertfordshire (UH) employs approximately 150 academic staff in a state-of-the-art environment located in Hatfield Business Park. It offers 17 undergraduate degree programmes and 21 postgraduate programmes; there are about 80 research students, mostly working at doctoral level. The University of Hertfordshire is the UK's leading business-facing university and an exemplar in the sector. It is one of the region's largest employers with over 2,300 staff and a turnover of £231million. In the 2008 UK Research Assessment Exercise it was given the highest rank for research quality among the post-1992 universities. It has a student community of over 27,000 including more than 2,900 international students. The University of Hertfordshire was awarded 'Entrepreneurial University of the Year 2010' by the Times Higher Education (THE) and ranks in the top 4% of all universities in the world according to the latest THE World University Rankings.

A DECISION SUPPORT TOOL FOR PREDICTING PATIENTS AT RISK OF READMISSION: A COMPARISON OF CLASSIFICATION TREES, LOGISTIC REGRESSION, GENERALIZED ADDITIVE MODELS, AND MULTIVARIATE ADAPTIVE REGRESSION SPLINES

Eren Demir

Department of Marketing & Enterprise, Business Analysis and Statistics Group,
The BusinessSchool, University of Hertfordshire, Hertfordshire, UK

Abstract

The number of emergency (or unplanned) readmissions in the United Kingdom National Health Service (NHS) has been rising for many years. This trend, which is possibly related to poor patient care, places financial pressures on hospitals and on national healthcare budgets. As a result, clinicians and key decision makers (e.g. managers and commissioners) are interested in predicting patients at high risk of readmission. Logistic regression is the most popular method of predicting patient-specific probabilities. However, these studies have produced conflicting results with poor prediction accuracies. We compared the predictive accuracy of logistic regression with that of regression trees for predicting emergency readmissions within forty five days after been discharged from hospital. We also examined the predictive ability of two other types of data-driven models: generalized additive models (GAMs) and multivariate adaptive regression splines (MARS). We used data on 963 patients readmitted to hospitals with chronic obstructive pulmonary disease. We used repeated split-sample validation: the data were divided into derivation and validation samples. Predictive models were estimated using the derivation sample and the predictive accuracy of the resultant model was assessed using a number of performance measures, such as area under the receiver operating characteristic (ROC) curve in the validation sample. This process was repeated 1000 times—the initial data set was divided into derivation and validation samples 1000 times, and the predictive accuracy of each method was assessed each time. The mean ROC curve area for the regression tree models in the 1000 derivation samples was 0.928, while the mean ROC curve area of a logistic regression model was 0.924. Our study shows that logistic regression model and regression trees had performance comparable to that of more flexible, data-driven models such as GAMs and MARS. Given that the models have produced excellent predictive accuracies, this could be a valuable decision support tool for clinicians (health care managers, policy makers, etc.) for informed decision making in the management of diseases, which ultimately contributes to improved measures for hospital performance management.

1. INTRODUCTION

Under new United Kingdom (U.K.) Government plans the National Health Service (NHS) hospitals will face financial penalties if patients are readmitted as an emergency within a short period of time after being discharged (Department of Health, 2010). Hospitals in England will be paid for initial treatment but not paid again if a patient is brought back in with a related problem. This raises serious concerns for clinicians and hospital managers, that is, how do you identify and prevent those patients at risk of readmission after discharge?

The ability to identify emerging risk patients will enable NHS organisations to take a more strategic approach to their care management interventions. For example, Clinical Commissioning Groups (CCGs) will be able to design and implement interventions and care pathways along the continuum of risk, ranging from, (i) prevention and wellness promotion for relatively low risk patients, (ii) supported self-care interventions for moderate risk patients, (iii) early intervention care management for patients with emerging risk, and (iv) intensive case management for very high risk patients. Therefore, accurately predicting will allow CCGs for effective patient risk stratification, thus permitting personalised care plan in the community for vulnerable patients most at risk. CCGs are groups of General Practitioners and from April 2013 they will be responsible for designing local health services in England.

There is a growing body of literature attempting to describe and validate hospital readmission risk prediction tools (Kansagara, Englander, Salanitro, Kagen, Theobald, Freeman & Kripalani, 2011). These models have been categorised into three:

(i) models relying on retrospective administrative data (Bottle, Aylin, & Majeed ,

2006; Krumholz, et al., 2008a; Krumholz, et al., 2008b; Krumholz, et al. 2008c; Hammill, et al., 2011; Howell, Coory, Martin, & Duckett, 2009; Holman, Preen, Baynham, Finn, & Semmens, 2005)

(ii) models using real-time administrative data (Amarasingham, et al., 2010; Billings & Mijanovich, 2007; Billings, Dixon, Mijanovich, & Wennberg , 2006), and

(iii) models incorporating primary data collection (e.g. survey or chart review data) (Coleman, Min, Chomiak, & Kramer, 2004; van Walraven, et al., 2010; Hasan, et al., 2010).

Most of these models had poor predictive ability, where the area under the receiver operating characteristic (ROC) curve ranged from 0.55 to 0.72, except Coleman et al. (2004) used administrative data on comorbidity and prior use of medical services along with functional status data and reported ROC curve value of 0.83. Here, ROC is defined to be the proportion of times the model correctly discriminates a pair of readmitted and non-readmitted patients. The area under the curve of 0.50 indicates that the model performs no better than chance; 0.70 to 0.80 indicates modest or acceptable discriminative ability; and a value of greater than 0.80 indicates good discriminative ability.

Logistic regression is the most commonly used method for predicting the probability of an adverse outcome in the medical literature (including the readmission prediction models mentioned above). Recently, data-driven methods, such as classification and regression trees (CART) have been used to identify subjects at risk of adverse outcomes or of increased risk of having specific diagnoses (Nishida, et al., 2005; Kuchibhatla & Fillenbaum, 2003; Avila, Segal, Wong, Boushey, & Fahy, 2000; Schwarzer, Nagata, Mattern, Schmelzeisen, & Schumacher, 2003; Hasford, Ansari, & Lehmann, 1993; Stewart & Stamm, 1991; Sauerbrei,

Madjar, & Prompeler, 1998; El-Solh, Sikka, & Ramadan, 2001; Tsien, Fraser, Long, & Kennedy, 1998; James, White, & Kraemer, 2005; Long, Griffith, Selker, & D'Agostino, 1993; Nelson, Bloch, Longstreth, & Shi, 1998; Lemon, Roy, Clark, Friedmann, & Rakowski, 2003). Decision rules generated by CART could easily be interpreted and applied in clinical practice. Furthermore, CART methods are versatile at identifying important interactions in the data and in identifying clinical subgroups of subjects at very high or very low risk of adverse outcomes (Lemon et al., 2003).

There are a number of studies that compared the performance of regression trees and logistic regression for predicting outcomes (Austin, 2007). Austin (2007) grouped these studies into three broad categories. First, studies that compared the significant predictors found by logistic regression with the variables identified by a regression tree analysis as predictors of the outcome. Second, studies that compared the sensitivity and specificity of logistic regression with that of regression trees. Third, studies that compared the predictive accuracy, as measured by the area under the ROC curve, of logistic regression with that of regression trees. Among these studies, the conclusions were found to be inconsistent. Six studies concluded that regression trees and logistic regression had comparable performance; five studies concluded that logistic regression had superior performance to regression trees.

The current study had two objectives: First, to compare the predictive ability of conventional logistic regression with that of regression tree methods for predicting patients at risk of readmission within 45 days after discharge. Second, to compare the relative performance of two other-data driven methods of analysis: generalized additive models (GAMs) and multivariate adaptive regression splines (MARS) models. We used repeated split sample validation using a large sample of patients hospitalised with chronic obstructive pulmonary

disease (COPD) and asthma in a primary care trust in England. COPD is known to be one of the leading causes of emergency readmission in the UK (Roland, Dusheiko, Gravelle, & Parker, 2005).

For the first time ever, four well established methodologies are rigorously evaluated and compared using newly derived variables (that have never being tested before) for the purpose of predicting patients at risk of readmission. Given that the NHS faces an unprecedented resource challenge: net savings of £20 billion must be achieved over the coming 3-4 years, representing a productivity challenge of around 4% a year (Hamm, 2010); the NHS policy documents stress the importance of measuring outcomes (i.e. patient readmissions) (Department of Health, 2011), and the Government's initiative towards expanding case management (i.e. managing patients in the community) (Department of Health, 2005), this research will make a timely contribution.

The development of models that are capable of accurately predicting patients at risk of readmission will inevitably enable clinicians, practitioners and senior managers' to improve clinical outcomes and increase effective budgeting. This research will also support positive patient centred outcome for the local population through more timely and effective and cost effective interventions (reduction in waiting lists, mitigating financial risks, hence cost saving). Therefore, this could be a valuable decision support tool for clinicians (health care managers, policy makers, etc.) for informed decision making in the management of diseases, which ultimately contributes to improved measures for hospital performance management.

2. METHODS

2.1 Data Sources

The data provided by a primary care trust (PCT) in England comprised three key sources of data: inpatient care, outpatient care, and accident & emergency (A&E). The inpatient care dataset provides a wide variety of information on admissions to NHS hospitals including patient details, when and where they were treated, care period, diagnosis, discharge, and geographical data. The Outpatient dataset contains information on outpatient appointments to NHS hospitals (day cases). It includes appointment dates, attendance types and non-attendances, waiting times, clinical and geographical data, patient details, socio economic factors, referral source and outcome results. The A&E dataset provides information on patient accident and emergencies to NHS hospitals including reason for and location of accident, hospital arrival, diagnosis, disposal, type of department attended, waiting times, and referral source. A full list of variables can be obtained from the hospital episodes statistics website (Hospital Episode Statistics, 2012).

The data was provided in Microsoft Access and Excel format and necessary steps were taken to import the data into MySQL version 5.0, so that database programming could be carried out to prepare the data for analysis. Initial checks were made to ensure that the data sets provided contained encrypted NHS numbers for matching purposes. The data period is from 01/04/08 to 31/12/10 (approximately 2.75 years). The total number of observations in the A&E dataset is 275,366 records, 122,446 inpatient care admissions, and 1,022,113 outpatient attendances. The first two years is used to develop the predictive models (derivation sample) and the third year (validation sample) to evaluate the observed vs. predicted results.

To model a representative subset of patients, we select patients according to either primary diagnosis or main specialty. We focus on patients with COPD and asthma as these are known to be the leading causes of early readmission in the UK (Roland et al., 2005). Readmission time is the time (in days) from the date of discharge to next emergency admission. Patients readmitted within 45 days after discharge (respectively, greater than 45 days) are classified as high risk group of readmission (respectively, low risk group), hence a binary response variable.

Patients who had the primary diagnosis codes corresponding to COPD (ICD-10 codes J40–J44) and asthma (ICD-10 code J45) were extracted. After data cleansing process (e.g. removing missing values and outliers) the total number of readmitted COPD and asthma patients during the 2.75 year period was 963 (413 and 550 in the high risk and low risk group of readmission, respectively). All variables listed in Table 1 were derived through database programming. The variables are categorised into three areas: medical comorbidity, prior use of medical services, patient characteristics, socio demographic and social determinants.

Around 14.8% of patients were diagnosed with two or more long term conditions (LTCs) (Table 1). A LTC is defined to be COPD, asthma, coronary artery disease, congestive heart failure, hypertension and cancer. Note that all explanatory variables were derived based on the admission date prior to readmission, for example, if a patient is readmitted on 01/06/2009, then the derived variables for this particular patient is based on the data prior to 01/06/2009. Approximately 22.7% of patients had one emergency readmission in the past 30 days, whereas 30.9% had three or more emergency admissions in the past 365 days. The average total previous length of stay prior to emergency admission in the last 30 days is around six days (25^{th} and 75^{th} percentile is 0 and 7 days, respectively). Interestingly, almost a fifth of

patients had 6-10 outpatient attendances in the past two years. Furthermore, fifty five per cent had three or more A&E visits in the past one year.

Insert Table 1 here

[All figures and tables are appended at the end of the paper.]

2.2. Initial Model for Predicting Patients at Risk of Readmission

A parsimonious model for predicting patients at risk of readmission was implemented based on the univariate logistic regression method using repeated bootstrap resampling approach. This model was developed by drawing repeated bootstrap samples from the sample of COPD and asthma patients. Those variables that were identified as significant predictors of risk of readmission in at least 75 per cent of the bootstrap samples were retained for inclusion in the final predictive model. The resultant model comprised thirty eight variables (listed in Table 2), that is, 38 were found to be significant in at least 75 per cent of the bootstrap samples out of the 79 variables listed in Table 1. This model will be used as the basis for some of the regression models that we will consider in this study.

Insert Table 2 here

2.3. Predictive Models for Risk of Readmission

In this section, we describe the four different classes of predictive models that were used to predict patients at risk of readmission. All model fitting and model validation was done using the R statistical programming language (R Core Development Team, 2005).

2.3.1. Logistic regression

Three separate logistic regression models were developed to predict patients at risk of readmission. The first model consisted of the thirty eight variables described in section 2.2, known as the reduced model. The second model was constructed using backwards variable elimination. This consisted of all the 38 variables along with all two-way interactions, with the thirty eight main effects being forced to remain in each model. The third model was also constructed using backwards variable elimination. However, in this instance, the initial model consisted of the 79 variables listed in Table 1. The logistic regression models were fitted using the **glm** function in R.

Backwards variable elimination was done using the **step** function in R. This implementation of backwards variable elimination is based upon sequentially eliminating variables from an initial model. At each step the variable is removed from the current model that results in the greatest reduction in the Akaike Information Criterion (AIC). The process of eliminating variables terminates either when a pre-specified boundary model is achieved or when no step will cause a further reduction in the AIC criterion (Hastie & Pregibon, 1993).

2.3.2. Classification trees

Binary recursive partitioning methods are rarely used to construct regression trees to predict patients at risk of readmission. The R implementation of regression tree only allows for binary partitions (or splits). In addition, the R implementation only allows for splits on individual variables and does not allow for splits on linear combinations of predictor variables. At each node, classification tree partitions the input variables into a set of homogeneous regions. The splits should divide the observations within a node so that the class types within a split are mostly of one kind (i.e. readmitted or not readmitted).

One advantage of this approach is that it does not require the parametric specification of the nature of the relationship between predictor variables and the outcome. In addition, the assumptions of linearity that are frequently made in linear and generalized linear models are not required for tree-based regression methods. Furthermore, tree-based methods are adept at identifying important interactions between predictor variables.

An initial tree was grown using all 79 candidate predictor variables listed in Table 1. Once the initial regression tree had been grown, the tree was pruned. A cross validation was used on the derivation data set to determine the optimal number of leaves on the tree (Faraway, 2006). Predictions were obtained on the validation data set using the pruned tree. The regression tree models were fit using the **rpart** function in the **rpart** package for R. The following code was used in R:

```
COPD.tree <- rpart(readmission ~ x.1 + x.2 + · · · + x.79, data=COPD.derivation)        (1)
```

Finally, the tree is pruned using the **prune** function with the cost complexity parameter (cp)

```
prune.COPD.tree<-prune(tree.derive, cp=0.01)        (2)
```

This final regression tree fit to the derivation sample was then used to obtain predictions for patients in the validation sample.

2.3.3. Generalized additive models

A generalized additive model (GAM) is an additive regression model of the form

$$y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \epsilon \qquad (3)$$

where the f_j are smooth arbitrary functions (e.g. lowess and smoothing splines) (Hastie & Tibshirani, 1990). Additive models are more flexible than the linear model, but still interpretable since the functions f_j can be plotted to give a sense of the marginal relationship between the predictors and the response. Categorical variables can be easily accommodated within the model using the usual regression approach. For example,

$$y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + Z_\gamma + \epsilon \quad (4)$$

where Z is the design matrix for the variables that will not be modelled additively, where some may be quantitative and others qualitative. The γ are the associated regression parameters.

In the current study, we considered three separate GAMs for predicting patients at risk of readmission. First, we considered the reduced model described above (variables listed in Table 2). Total length of stay variables and age at admission were modelled using smoothing splines. A second model was fitted that consisted of the above GAM, along with all two-way interactions. The third model contained all 79 variables, while the 12 continuous variables were modelled using smoothing splines.

The GAMs were fitted using the **gam** function in the MGCV package in R. The first generalized additive was fit using the following code in R:

```
COPD.gam <- gam(readmission ~ s(x.1) + s(x.2) + · · · +s(x.9) +  x.10 + x.11 + · · ·  x.38, family
= binomial, data=COPD.derivation)                                                    (5)
```

Here, s(x.1) – s(x.9) are the continuous variables that were modelled using smoothing splines and x.11 – x.38 are the categorical variables listed in Table 2. This particular GAM is expressed as in equation (4).

2.3.4. Multivariate adaptive regression spline models

Multivariate adaptive regression splines (MARS) is an adaptive regression procedure well suited to problems with a large number of predictor variables (Friedman, 1991; Hastie, Tibshirani, & Friedman, 2001). The basic principle of MARS is that it divides the data into several regions, and fits a regression model to each region. MARS uses an expansion based on linear spline functions. For a given predictor X_j and a given value c taken by the predictor variable, one can define two linear spline functions: $(X_j - c)_+$ and $(c - X_j)_+$, where '+' refers to the positive part. For example, suppose X_j is 'age at admission' and the best split is at age 55 (i.e. $c = 55$), then $(X_j - 55)_+$ and $(55 - X_j)_+$ refers to the region greater and lower than 55, respectively.

We examined three separate MARS models. Each used the 79 variables described in Table 1. The first model was an additive model that did not allow interactions between the predictor variables. The second model allowed for the inclusion of two-way interactions, while the third model allowed for the inclusion of all possible interactions, including 79-way interactions. MARS models are constructed using generalized cross-validation to determine the optimal number of terms in the model (Hastie, Tibshirani, & Friedman, 2001). This use of generalized cross-validation helps protect against over-fitting the model in the derivation sample. Thus, while the third MARS model allowed for the potential inclusion of all possible

interactions, the use of generalized cross-validation minimizes the likelihood that the final model will be over-fit to the derivation sample. The MARS models were fit using the **earth** function in the EARTH package. The following code was used to fit the first MARS model in R:

```
COPD.mars <- mars(readmission ~ x.1 + x.2 + … + x.79, data = COPD.derivation, degree=1,
                  glm=list(family=binomial)))                                    (6)
```

where COPD.derivation denotes a matrix of the predictor variables (from the derivation sample), readmission denotes the binary outcome variable and degree = 1 refers to additive model (i.e. a model with no interactions).

2.4. Evaluating the predictive powers of models

Motivated by Austin (2007), repeated split-sample validation was used to compare the predictive accuracy of each statistical method. The data were divided into derivation and validation components. The first two years of data were used for model derivation and the remaining nine months of data was used for model validation. Each derivation sample consisted of 725 patients, while each validation sample consisted of 238 patients. This process was repeated 1000 times.

Each model was fitted on the derivation sample. Predictions were then obtained for each patient in the validation sample using the model derived on the derivation sample. The predictive accuracy of each model was summarized by the area under the ROC curve, which is equivalent to the c-statistic (Harrell Jr, 2001). In medical literature, ROC is the most widely used statistic to assess the predictive power of models related to predicting adverse events. The area under the ROC curve was obtained for both the derivation and validation samples.

In addition to ROC, Austin (2007) used a number of other measures (suggested by Harrell Jr (2001)), such as the generalized R_N^2 index (Nagelkerke, 1991) and Brier's score (Harrell Jr, 2001). Brier's score is defined as

$$B = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{P}_i - Y_i\right)^2 \quad (7)$$

where $\hat{P}_i$ is the predicted probability and Y_i is the observed response for the i th patient. We computed the generalized R_N^2 index and Brier's score in each of the validation samples. The area under the ROC curve, the generalized R_N^2 index, and Brier's score were computed using the **val.prob** function from the RMS package for R. We also report the sensitivity score and specificity score for the validation samples. Sensitivity score is the proportion of correctly classified readmitted patients, whereas specificity score is the proportion of correctly classified non-readmitted cases. These two measures are well known by clinical and managerial staff in the National Health Service.

The above methods were repeated 1000 times: the initial data were divided into derivation and validation components 1000 times. Each predictive model was fitted using the derivation data set and predictions were then obtained on the validation data set. Results were then summarized over the 1000 validation data sets. By using 1000 different derivation/validation samples, we were able to assess the robustness of our results under different derivation and validation samples. This process was carried out using the **boot** function from the BOOT package for R.

3. RESULTS

3.1. Predictive Performance

The mean area under the ROC curve for each model in both derivation and validation samples are reported in Table 3.

Insert Table 3 here

In the validation sample, the mean ROC curve area for the regression tree model was 0.924, while the mean ROC curve area for the stepwise logistic regression model was 0.928. The difference in ROC curve areas for the regression tree method and the stepwise logistic regression model ranged from a low of zero to a high of 0.137 across the 1000 validation samples (mean difference: 0.026). The ROC curve areas for the other modelling methods in the validation samples ranged from a low of 0.824 (MARS model with all interactions) to a high of 0.924 (GAM and MARS with all variables, i.e., full model). As both models included all variables (i.e. 79) this may mean that the full models were over fitted on the derivation samples.

The mean ROC curve area for the regression tree decreased from 0.948 in the derivation samples to 0.924 in the validation samples – a very small decrease of 0.024. The decline in the mean ROC curve area between the derivation and validation samples was negligible for the logistic regression with the backwards elimination from the full model (0.049). Similarly, the drop in ROC curve area from the derivation sample to the validation sample was identical and small for the GAM (full model: 0.978 to 0.924) and the MARS full model. The highest difference in ROC curve areas between the derivation and validation samples was observed to be in the MARS model with all interactions (0.171), followed by MARS with two-way interactions (0.131) and the logistic regression with two-way interactions (0.122). The greater

decline in ROC curve area for the MARS models compared to the simpler logistic regression and regression tree may be indicative of a tendency of the more complex MARS models to over-fit on the derivation samples. For most models, the decrease in ROC curve area from the derivation sample to the validation sample was relatively modest.

The distribution of the area under the ROC curve in the 1000 validation data sets for each modelling approach is described in Figure 1. The distribution of ROC curve areas for the MARS model with two-way interactions and all interactions, logistic regression with two-way interactions, and GAM with two-way interactions shifted downwards (and to the left) compared to that of the other modelling approaches (i.e. greater variability in ROC curve areas). This clearly demonstrates that models with two way interactions (or higher) had consistently poor performance than the other models.

The distributions of ROC curve areas for the logistic regression model obtained from the full model using backwards elimination, the reduced logistic regression model, and the regression tree model were almost identical. The same phenomenon was observed for the GAM reduced model and GAM full model.

Insert Figure 1 here

The generalized R_N^2 index is reported in Table 3 for each of the modelling strategies. The index ranged from a low of 0.701 for MARS with all interactions (and logistic regression with two-way interactions) to a high of 0.859 for the logistic regression model obtained from the full model using backwards elimination, that is, this model explained substantial proportion of the observed variation. The Brier's score is also reported in Table 3 for each of the modelling strategies. Of note is the fact that there isn't too much variability in the

estimated Brier's score for all the modelling strategies except logistic regression and GAM with two-way interactions and MARS with all interactions. The distribution of the generalized R_N^2 index and Brier's scores in the 1000 validation data sets for each modelling approach are described in Figures 2 and 3, respectively.

Insert Figure 2 here

Insert Figure 3 here

Similar observations can be made concerning the distribution of the generalized R_N^2 index in the validation samples as was made above for the distribution of the ROC curve areas for the different models in the validation samples. In the case of the distribution of Brier's score (Figure 3) the logistic regression model (two-way interactions) shifted the most to the right (i.e. the least effective model), whereas the regression tree, logistic regression (obtained from backwards elimination), and GAM (full model) exhibited the least variability with the most effective predictive ability. The remaining models can be considered to be comparable (0.137 to 0.151). Sensitivity and specificity scores are also reported in Table 3. The highest mean sensitivity and specificity scores over the 1000 validation samples is from the regression tree model (0.862 and 0.904, respectively) and logistic regression based on backwards elimination from the full model (0.821 and 0.897, respectively). Figures 4 and 5 illustrate the distribution of sensitivity and specificity scores, respectively. Note that the distribution of specificity scores for all modelling strategies (except logistic regression with two-way interactions) are very similar, whereas some variability is observed in the distribution of sensitivity scores.

Insert Figure 4 here

Insert Figure 5 here

3.2.Miscellaneous results

The best modelling strategy is selected from each method and key findings are presented here. Judging from the predictive performance measures from Table 3, the stepwise logistic regression, GAM (full model) and MARS (full model) are selected. The regression tree obtained using one of the derivation samples is illustrated in Figure 6.

Insert Figure 6 here

This particular regression tree had seven terminal nodes. Six variables were used in creating the tree which are all related to prior use of medical services (e.g. length of stay, previous history of readmissions). If a patient had one previous emergency readmission in the last 30 days, then there is a 100% chance of being readmitted within 45 days after discharge (N=166). This can be considered to be clinically relevant, as patient's short previous history is related to risk of adverse outcomes (e.g. readmission, mortality). Similarly, there is a 90% chance of a patient being readmitted if their previous length of stay in hospital (emergency and non-emergency) was greater than half a day and had two or more emergency admissions in the past 90 days (N = 80).

From the logistic regression obtained from backwards elimination, twenty three variables (out of a total of 79) were included in the final model, of which nineteen had a p-value less than 0.05. Patients who had four or more distinct in-patient primary diagnosis (i.e. clinical conditions) were eight times more likely to have been readmitted. Interestingly, those who had two or more emergency readmissions in the last 90 days were 65 times more likely to be readmitted.

Figure 7 describes the relationship between log odds of readmission and "total emergency and non-emergency length of stay in the past 90 days" and "total emergency length of stay in the past 90 days" from the GAMs full model. Note that these two variables are not the same, where the latter takes into account length of stays based on emergency admissions only, whereas the former includes non-emergency admissions as well (e.g. planned surgeries). For each value of a given variable, we determined the predicted log-odds of readmission, holding the other continuous variables fixed at the sample average and the binary predictor variables set to absent.

Insert Figure 7 here

One observes that the relationship between the two variables and log-odds of readmission is non-linear up-to 30 days of length of stay and approximately linear thereafter. The risk of readmission increases with an increasing total emergency length of stay in the past 90 days prior to next readmission. Conversely, increasing LoS at emergency and non-emergency admissions reduces the risk of being readmitted. One explanation for this phenomenon is that when "emergency" admissions are coupled with "non-emergency admissions", further complications at the non-emergency admission phase for sick and frail patients may have been prevented, and consequently the more the patient is cared/treated at both admissions, the lower the risk of future readmissions.

Figure 8 describes the relationship between the log-odds of readmission and the two LoS variables used above from MARS full model. One observes that up-to a total of 10 LoS days (emergency and non-emergency LoS) log odds of readmission increase rapidly, where this risk gradually decreases after a LoS greater than 20 days. In relation to emergency LoS only,

the log-odds of readmission are low for patients who stayed in hospital for less than six days (in the last 90 days prior to next readmission) and gradually increases thereafter.

Rapid patient discharge to free beds for incoming patients is a controversial debate in the UK. Some argue that patients may have been discharged too soon, raising the issue that patients are being discharged 'sicker and quicker' (Capewell, 1996). As a result, early discharges may generate high levels of readmissions, which could possibly be seen as patients being discharged inappropriately. In this respect, Figures 7 and 8 points to the fact that when patients are cared and treated for longer periods of time (as an emergency and non-emergency LoS) the risk of future readmission diminishes, which may result in the reduction of emergency readmissions, increase patient and staff satisfaction, reduce waiting lists, increase the performance of the hospital, and given the economic conditions in the UK, cost savings.

Insert Figure 8 here

4. DISCUSSION

In the National Health Service, changes to commissioning arrangements have increased the focus and drive to reduce hospital admissions. Approximately 35% of hospital admissions in England are emergency admissions costing £11 billion per annum (2010/11), which represent 36.7% of hospital admissions in England (5.3 million admissions in 2010/11). Given that the tough economic conditions are expected to be with us for quite a while in the future, the UK Government's target is to provide personalised care plan for vulnerable people most at risk. Managing emergency readmissions will inevitably reduce the burden on non-emergency health care and resource use, which may lead to substantial amount of cost savings, reduction on waiting lists, and more importantly positive patient centred outcomes for patients and

carers. Therefore, an appropriate toolkit is needed to aid clinical commission groups in their intervention policies to provide treatment in the community to those patients who are at high risk of readmission. There has been an increasing interest in developing statistical models to identify patients at increased risk of readmission within a short period of time after discharge (e.g. 45 days). Many models have been developed in the UK and other countries where these studies produced conflicting findings, resulting in poor predictions.

In the current study, we have demonstrated, using a large sample of patients with chronic obstructive pulmonary disease that conventional logistic regression and regression trees produced comparable results to that of modern flexible regression methods such as GAMs and MARS models. The mean ROC curve area for conventional logistic regression with no interactions and regression trees was 0.928 and 0.924, respectively in the validation sample, while the corresponding value for GAMs and MARS was 0.854 and 0.721, respectively. The highest observed ROC curve area was for the logistic regression model obtained from the full model using backwards elimination. In addition to comparing the predictive accuracy of regression methods from different families of methods, we also compared the predictive accuracy of models with differing complexity from the same family of models. We found that more complex models from the same family had lower predictive accuracy (GAMs and MARS). Similar results were observed when the generalized R_N^2 index, Brier's score, sensitivity score and specificity score were used to quantify the predictive accuracy of different regression models.

Nevertheless, GAMs and MARS provided very useful insights. First, analyses conducted using GAMs indicated that the relationship between log-odds of readmission (within 45 days after discharge) and "total emergency length of stays in the past 90 days" was non-linear up-

to a LoS of 30 days and approximately linear thereafter. One can observe that the risk of readmission increases after patients having spent above 30 days as an emergency admission. Clinicians, nurses and key decision makers for COPD patients could pay particular attention to those patients who have been in emergency care for a total of 30 or more days. Note that thirty or more days refer to the cumulative length of stays in the past 90 days as emergency admissions only. This finding was also confirmed by the MARS full model (additive model).

The conventional logistic regression model was able to exploit the strong underlying linear relationships in the data. For example, commissioning managers (and the clinical team) would need to be extra cautious on patients who had two or more emergency admissions in the past 90 days prior to next readmission, simply due to the fact that this group of patients are 65 times more likely to be readmitted. The regression tree model partitioned the sample using binary decision rules and one useful partition was that if a patient had one previous emergency readmission in the last 30 days, then there is a 100% chance of being readmitted again. A closer look at this particular node revealed that out of the 166 patients that were assigned to this node, approximately 88% of patients were correctly identified in the validation sample.

We offer the following three suggestions to researchers and practitioners assessing the predictive accuracy of regression models to predict patients at risk of readmission. First, the data should be split into a derivation sample and validation sample, so that the predictive accuracy of regression models can be assessed using a summary measure such as the area under the ROC curve. Second, do not just rely on sensitivity and specificity scores (the proportion of correctly identified readmitted and non-readmitted cases, respectively), as this approach has been criticised for a variety of reasons (Harrell Jr, 2001). Third, repeated split-

sample validation should be employed to assess the variability in the performance measures across the 1000 validation samples. To the best of our knowledge, no study has ever compared CART and logistic regression with other data driven methods (GAMs and MARS) using repeated split sample validation approach to examine the robustness of the findings to predict patients at risk of readmission within forty five days after discharge.

In conclusion, we demonstrated that logistic regression had superior predictive ability compared to modern data-driven methods. Furthermore, regression trees had comparable predictive ability to the conventional logistic regression. A message to key decision makers in the NHS (and other countries) is that to the best of our knowledge this particular research has produced the highest predictive accuracies that have ever been published. Based on the rigorous evaluation of 1000 validation samples, the area under the ROC curve is 0.93 with an overall classification accuracy of 0.90 (the highest previously reported area under the ROC curve is 0.83 by Kansagara, et al., 2011). Therefore, the methods outlined in this study will enable practitioners and managers in the NHS to develop a robust decision support toolkit to provide treatment in the community to patients at high risk of readmission. This can be a valuable tool in helping to tailor community care to local needs and ultimately contribute to improved measures in reducing readmissions.

References

Amarasingham, R., Moore, B. J., Tabak, Y. P., Drazner, M. H., Clark, C. A., Zhang, S., . . . Halm, E. A. (2010). An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Medical Care, 48*(11), 981-988.

Austin, P. C. (2007). A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Statistics in Medicine, 26*, 2937-2957.

Avila, P. C., Segal, M. R., Wong, H. H., Boushey, H. A., & Fahy, J. V. (2000). Predictors of late asthmatic response. Logistic regression and classification tree analyses. *American Journal of Respiratory and Critical Care Medicine, 161*, 2092-2095.

Billings, J., & Mijanovich, T. (2007). Improving the management of care for high-cost Medicaid

patients. *Health Affairs (Millwood), 26*(6), 1643-1654.
Billings, J., Dixon, J., Mijanovich, T., & Wennberg , D. (2006). Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. *British Medical Journal, 333*(7563), 327.
Bottle, A., Aylin, P., & Majeed , A. (2006). Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis. *Journal of the Royal Society of Medicine, 99*(8), 406-414.
Capewell, S. (1996). The continuing rise in emergency admissions. *British Medical Journal, 312*, 991–992.
Coleman, E. A., Min, S. J., Chomiak, A., & Kramer, A. M. (2004). Post hospital care transitions: patterns, complications, and risk identification. *Health Services Journal, 39*(5), 1449-1465.
Department of Health. (2005). Case management competences framework for the care of people with long term conditions. Retrieved August 23, 2012, from http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4118101
Department of Health. (2010). *White Paper: Equity and excellence - Liberating the NHS.* Retrieved August 22, 2012, from http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/@ps/documents/digitalasset/dh_117794.pdf
Department of Health. (2011). Health and Social Care Bill. Retrieved August 23, 2012, from http://www.publications.parliament.uk/pa/cm201011/cmbills/132/11132.pdf
El-Solh, A. A., Sikka, P., & Ramadan, F. (2001). Outcome of older patients with severe pneumonia predicted by recursive partitioning. *Journal of the American Geriatrics Society, 49*, 1614-1621.
Faraway, J. J. (2006). *Extending the linear model with R: Generalized Linear, mixed effects and nonparametric regression models.* Chapman & Hall.
Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics, 19*, 1-67.
Hamm, C. (2010). The coalition government's plans for the NHS in England. *British Medical Journal, 341*, c3790.
Hammill, B. G., Curtis, L. H., Fonarow, G. C., Heidenreich, P. A., Yancy, C. W., Peterson, E. D., & Hernandez, A. F. (2011). Incremental Value of Clinical Data Beyond Claims to Predict 30-Day Outcomes Following Heart Failure Admission. *Cardiovascular Outcomes Research, 4*(1), 60-67.
Harrell Jr, F. E. (2001). *Regression Modeling Strategies.* New York: Springer.
Hasan, O., Meltzer, D. O., Shaykevich , S. A., Bell, C. M., Kaboli, P. J., Auerbach, A. D., . . . Schnipper, J. L. (2010). Hospital readmission in general medicine patients: a prediction model. *Journal of General Internal Medicine, 25*(3), 211-219.
Hasford, J., Ansari, H., & Lehmann, K. (1993). CART and logistic regression analyses of risk factors for first dose hypotension by an ACE-inhibitor. *Therapie, 48*, 479-482.
Hastie, T. J., & Pregibon, D. (1993). *Generalized linear models.* New York: Chapman & Hall.
Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized Additive Models.* London: Chapman & Hall.
Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* New York: Springer.
Holman, C. A., Preen, D. B., Baynham, N. J., Finn, J. C., & Semmens, J. B. (2005). A mutipurpose comorbidity scoring system performed better than the Charlson index. *Journal of Clinical Epidemiology, 58*(10), 1006-1014.
Hospital Episode Statistics. (2012). *HES online: Understanding the Data.* Retrieved August 24, 2012, from http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=289
Howell, S., Coory, M., Martin, J., & Duckett, S. (2009). Using routine inpatient data to indentify

patients at risk of hospital readmission. *BMC Health Services Research, 9*(96).
James, K. E., White, R. F., & Kraemer, H. C. (2005). Repeated split sample validation to assess logistic regression and recursive partitioning: an application to the prediction of congnitive impairment. *Statistics in Medicine, 24*, 3019-3035.
Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). Risk Prediction Models for Hospital Readmission. *Journal of the American Medical Association, 306*(15), 1688-1698.
Krumholz, H. M., Normand, S.-L. T., Keenan, P. S., Desai, M. M., Lin, Z., Drye, E. E., . . . Schreiner, G. B. (2008a). *Hospital 30-Day Acute Myocardial Infarction Readmission Measure: Methodology.* Retrieved August 20, 2012, from http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1219069855841
Krumholz, H. M., Normand, S.-L. T., Keenan, P. S., Desai, M. M., Lin, Z., Drye, E. E., . . . Schreiner, G. C. (2008b). *Hospital 30-Day Acute Myocardial Infarction Readmission Measure: Methodology.* Retrieved August 20, 2012, from http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1219069855841
Krumholz, H. M., Normand, S.-L. T., Keenan, P. S., Desai, M. M., Lin, Z., Drye, E. E., . . . Schreiner, G. C. (2008c). *Hospital 30-Day Pneumonia Readmission Measure.* Retrieved August 20, 2012, from http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1219069855841
Kuchibhatla, M., & Fillenbaum, G. G. (2003). Alternative statistical approaches to identifying dementia in a community-dwelling sample. *Aging and Mental Health, 7*, 383-389.
Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of Behavioral Medicine, 26*, 172-181.
Long, W. J., Griffith, J. L., Selker, H. P., & D'Agostino, R. B. (1993). A comparison of logistic regression to decision tree induction in a medical domain. *Computers and Biomedical Research, 26*, 74-97.
Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*, 691–692.
Nelson, L. M., Bloch, D. A., Longstreth, J. W., & Shi, H. (1998). Recursive partitioning for the identification of disease risk subgroups: a case-control study of subarachnoid hemorrhage. *Journal of Clinical Epidemiology, 51*, 199-209.
Nishida, N., Tanaka, M., Hayashi, N., Nagata, H., Takeshita, T., Nakayama, K., . . . Shizukuishi, S. (2005). Determination of smoking and obesity as periodontitis risks using the classification and regression tree method. *Journal of Periodontology, 76*, 923-928.
R Core Development Team. (2005). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.
Roland, M., Dusheiko, M., Gravelle, H., & Parker, S. (2005). Follow up of people aged 65 and over with a history of emergency admissions: Analysis of routine admission data. *British Medical Journal, 330*, 289-292.
Sauerbrei, W., Madjar, H., & Prompeler, H. J. (1998). Differentiation of benign and malignant breast tumors by logistic regression and a classification tree using Doppler flow signals. *Methods of Information in Medicine, 37*, 226-234.
Schwarzer, G., Nagata, T., Mattern, D., Schmelzeisen, R., & Schumacher, M. (2003). Comparison of fuzzy inference, logistic regression, and classification trees (CART). Prediction of cervical lymph node metastasis in carcinoma of the tongue. *Methods of Information in Medicine, 42*, 572-577.

Stewart, P. W., & Stamm, J. W. (1991). Classification tree prediction models for dental caries from clinical, microbiological, and interview data. *Journal of Dental Research, 70*, 1239-1251.
Tsien, C. L., Fraser, H. S., Long, W. J., & Kennedy, R. L. (1998). Using classification tree and logistic regression methods to diagnose myocardial infarction. *Medinfo, 9*, 493-497.
van Walraven, C., Dhalla, I. A., Bell, C., Etchells, E., Stiell, I. G., Zarnke, K., . . . Forster, A. (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Canadian Medical Association Journal, 182*(6), 551-557.

Table 1: Derived variables and characteristics of the study sample. 0/1 refers to dichotomous variables with their corresponding proportion of cases. For example, 14.8% of COPD and asthma patients have two or more long term conditions (LTCs). The index of multiple deprivation (IMD) score is a weighted index based on seven factors: income, employment, health and disability, education, skills and training, barriers to housing and services, living environment and crime.

Medical comorbidity	
One and only one long term condition (0/1)	54.1 per cent
Two or more long term conditions (0/1)	14.8 per cent
Two distinct in-patient primary diagnosis (0/1)	22.6 per cent
Three distinct in-patient primary diagnosis (0/1)	10.4 per cent
Four and above distinct in-patient primary diagnosis	22.2 per cent
Prior use of medical services: Inpatient care	
One emergency admission in the past 30 days (0/1)	22.7 per cent
More than one emergency admission in the past 30 days (0/1)	5.4 per cent
One emergency admission in the past 90 days (0/1)	32.5 per cent
Two or more emergency admissions in the past 90 days (0/1)	19.1 per cent
One emergency admission in the past 180 days (0/1)	33.2 per cent
Two or more emergency admissions in the past 180 days (0/1)	33.7 per cent
One emergency admission in the past 365 days (0/1)	35.1 per cent
Two emergency admissions in the past 365 days (0/1)	14.7 per cent
Three or more emergency admissions in the past 365 days (0/1)	30.9 per cent
One emergency admission in the past 730 days (0/1)	35.1 per cent
Two emergency admissions in the past 730 days (0/1)	15.8 per cent
Three emergency admissions in the past 730 days (0/1)	9.0 per cent
Four or more emergency admissions in the past 730 days (0/1)	28.2 per cent
Total previous emergency length of stay prior to emergency admission in the last 30 days	0.9 (0-0)
Total previous emergency length of stay prior to emergency admission in the last 90 days	3.3 (0-3)
Total previous emergency length of stay prior to emergency admission in the last 180 days	6.0 (0-7)
Total previous emergency length of stay prior to emergency admission in the last 365 days	10.2 (0-12)
Total previous emergency length of stay prior to emergency admission in the last 730 days	14.6 (1-16)
Total previous emergency and non-emergency length of stay prior to emergency admission in the last 30 days	1.0 (0-0)
Total previous emergency and non-emergency length of stay prior to emergency admission in the last 90 days	3.5 (0-4)
Total previous emergency and non-emergency length of stay prior to emergency admission in the last 180 days	6.4 (0-8)
Total previous emergency and non-emergency length of stay prior to emergency admission in the last 365 days	10.8 (0-13)
Total previous emergency and non-emergency length of stay prior to emergency admission in the last 730 days	15.3 (1-17)
One previous emergency readmission as high risk group in the last 730 days (0/1)	41.7 per cent
Two previous emergency readmission as high risk group in the last 730 days (0/1)	24.8 per cent
Three previous emergency readmission as high risk group in the last 730 days (0/1)	17.4 per cent
Four previous emergency readmission as high risk group in the last 730 days (0/1)	12.7 per cent
Five previous emergency readmission as high risk group in the last 730 days (0/1)	9.7 per cent
Prior use of medical services: Outpatient care	

One out-patient specialty visit in the last 30 days (0/1)	15.9 per cent
Two out-patient specialty visit in the last 30 days (0/1)	4.4 per cent
Three or more out-patient specialty visit in the last 30 days (0/1)	2.4 per cent
One out-patient specialty visit in the last 90 days (0/1)	20.8 per cent
Two out-patient specialty visit in the last 90 days (0/1)	11.4 per cent
Three or more out-patient specialty visit in the last 90 days (0/1)	13.9 per cent
1-5 out-patient specialty visits in the last 730 days (0/1)	38.7 per cent
6-10 out-patient specialty visits in the last 730 days (0/1)	19.7 per cent
Eleven or more out-patient specialty visits in the last 730 days (0/1)	15.2 per cent
Prior use of medical services: Accident & Emergency	
If the patient had an X-ray in their A&E visit in the last 180 days (0/1)	53.0 per cent
Arrived by ambulance in the last 90 days (0/1)	41.2 per cent
The patient was discharge to hospital in the last 180 days (0/1)	53.1 per cent
One A&E visit in the last 365 days (0/1)	16.1 per cent
Two A&E visit in the last 365 days (0/1)	14.5 per cent
Three or more A&E visit in the last 365 days (0/1)	55.2 per cent
Patient characteristics, socio demographic and social determinants	
Age group 0-4 (0/1)	6.0 per cent
Age group 5-14 (0/1)	4.5 per cent
Age group 15-39 (0/1)	9.3 per cent
Age group 40-59 (0/1)	18.5 per cent
Age group 60-64 (0/1)	6.7 per cent
Age group 65-69 (0/1)	8.4 per cent
Age group 70-74 (0/1)	13.2 per cent
Age group 75-79 (0/1)	10.5 per cent
Age group 80-84 (0/1)	14.1 per cent
Age group 85-89 (0/1)	6.5 per cent
Age group 90-94 (0/1)	2.2 per cent
Age 95+ (0/1)	0.1 per cent
Gender (female) (0/1)	51.2 per cent
Age (continuous variable)	60 (51-79)
Index of multiple deprivation (continuous variable)	24.9 (18-31)
Ethnicity	
British (White) (0/1)	63.9 per cent
Irish (White) (0/1)	3.3 per cent
Any other white background (0/1)	1.3 per cent
White and Black Caribbean (Mixed) (0/1)	0.4 per cent
White and Black African (Mixed) (0/1)	0.3 per cent
White and Asian (Mixed) (0/1)	0.2 per cent
Indian (Asian or Asian British) (0/1)	14.1 per cent
Pakistani (Asian or Asian British) (0/1)	3.3 per cent
Bangladeshi (Asian or Asian British) (0/1)	0.1 per cent
Any other Asian background (0/1)	3.7 per cent
Caribbean (Black or Black British) (0/1)	0.3 per cent
African (Black or Black British) (0/1)	0.8 per cent
Any other Black background (0/1)	0.8 per cent
Chinese (other ethnic group) (0/1)	0.2 per cent
Any other ethnic group (0/1)	7.1 per cent

Table 2: List of variables for the reduced model

One and only one long term condition (0/1)
Two or more long term conditions (0/1)
Two distinct in-patient primary diagnosis (0/1)
Three distinct in-patient primary diagnosis (0/1)
Four and above distinct in-patient primary diagnosis
One emergency admission in the past 90 days (0/1)
Two or more emergency admissions in the past 90 days (0/1)
One emergency admission in the past 180 days (0/1)
Two or more emergency admissions in the past 180 days (0/1)
Three or more emergency admissions in the past 365 days (0/1)
One emergency admission in the past 730 days (0/1)
Two emergency admissions in the past 730 days (0/1)
Three emergency admissions in the past 730 days (0/1)
Four or more emergency admissions in the past 730 days (0/1)
Total previous emergency length of stay prior to emergency admission in the last 90 days
Total previous emergency length of stay prior to emergency admission in the last 180 days
Total previous emergency length of stay prior to emergency admission in the last 365 days
Total previous emergency length of stay prior to emergency admission in the last 730 days
Total previous emergency and non-emergency length of stay prior to emergency admission in the last 90 days
Total previous emergency and non-emergency length of stay prior to emergency admission in the last 180 days
Total previous emergency and non-emergency length of stay prior to emergency admission in the last 365 days
Total previous emergency and non-emergency length of stay prior to emergency admission in the last 730 days
One previous emergency readmission as high risk group in the last 730 days (0/1)
Two previous emergency readmission as high risk group in the last 730 days (0/1)
Three previous emergency readmission as high risk group in the last 730 days (0/1)
Four previous emergency readmission as high risk group in the last 730 days (0/1)
Five previous emergency readmission as high risk group in the last 730 days (0/1)
One out-patient specialty visit in the last 30 days (0/1)
1-5 out-patient specialty visits in the last 730 days (0/1)
Eleven or more out-patient specialty visits in the last 730 days (0/1)
If the patient had an X-ray in their A&E visit in the last 180 days (0/1)
Arrived by ambulance in the last 90 days (0/1)
The patient was discharge to hospital in the last 180 days (0/1)
One A&E visit in the last 365 days (0/1)
Two A&E visit in the last 365 days (0/1)
Three or more A&E visit in the last 365 days (0/1)
Age (continuous variable)
If the patient had a long term condition in the past (0/1)

Table 3: Model discrimination in the 1000 repeated split samples

Model	ROC area: derivation sample	ROC area: validation sample	R_N^2: validation sample	Brier's score: validation sample	Sensitivity score: validation sample	Specificity score: validation sample
Regression Tree	0.948	0.924	0.721	0.089	0.862	0.904
Logistic regression (backwards elimination from full model)	0.977	0.928	0.859	0.101	0.821	0.897
Logistic regression (reduced model)	0.954	0.880	0.759	0.137	0.773	0.868
Logistic regression (two-way interactions)	0.976	0.854	0.701	0.198	0.701	0.804
GAM (full model)	0.978	0.924	0.854	0.106	0.802	0.896
GAM (reduced model)	0.959	0.875	0.778	0.138	0.753	0.863
GAM (two-way interactions)	0.963	0.856	0.792	0.149	0.744	0.851
MARS (full model)	0.978	0.924	0.721	0.142	0.813	0.892
MARS (two-way interactions)	0.991	0.863	0.712	0.143	0.794	0.879
MARS (all interactions)	0.995	0.824	0.701	0.151	0.801	0.872

Figure 1: Distribution of ROC curve areas in 1000 validation samples

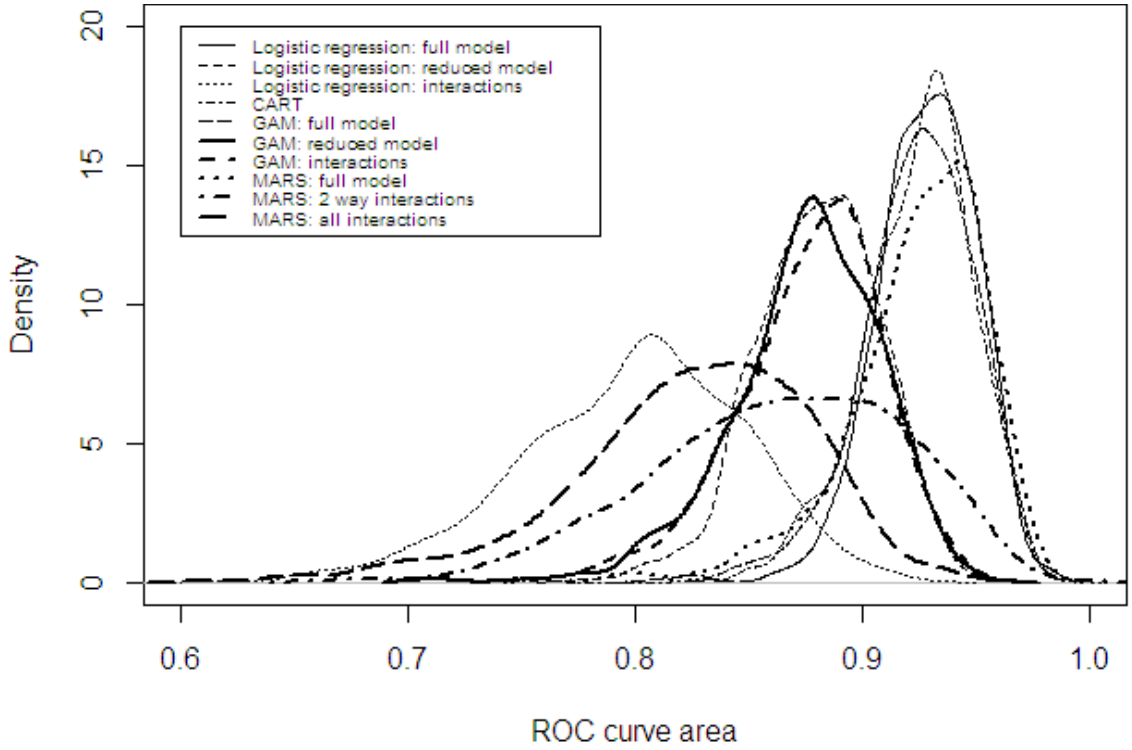

Figure 2: Distribution of R_N^2 index in 1000 validation samples

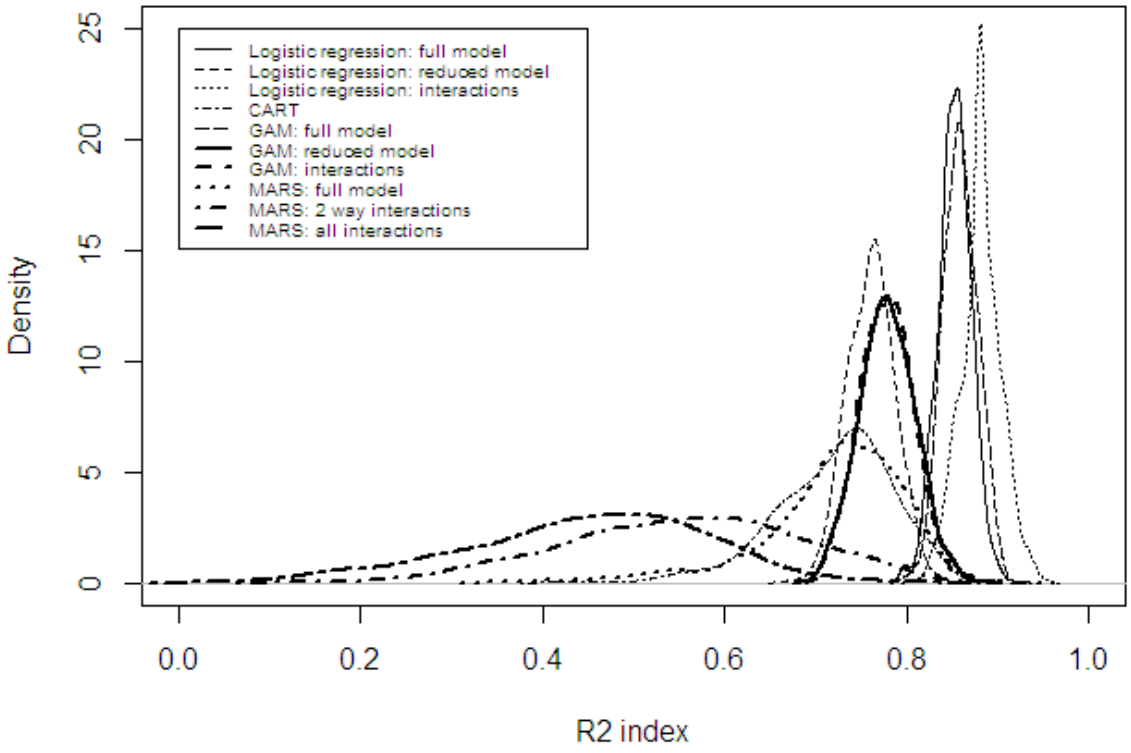

Figure 3: Distribution of Brier's scores in 1000 validation samples

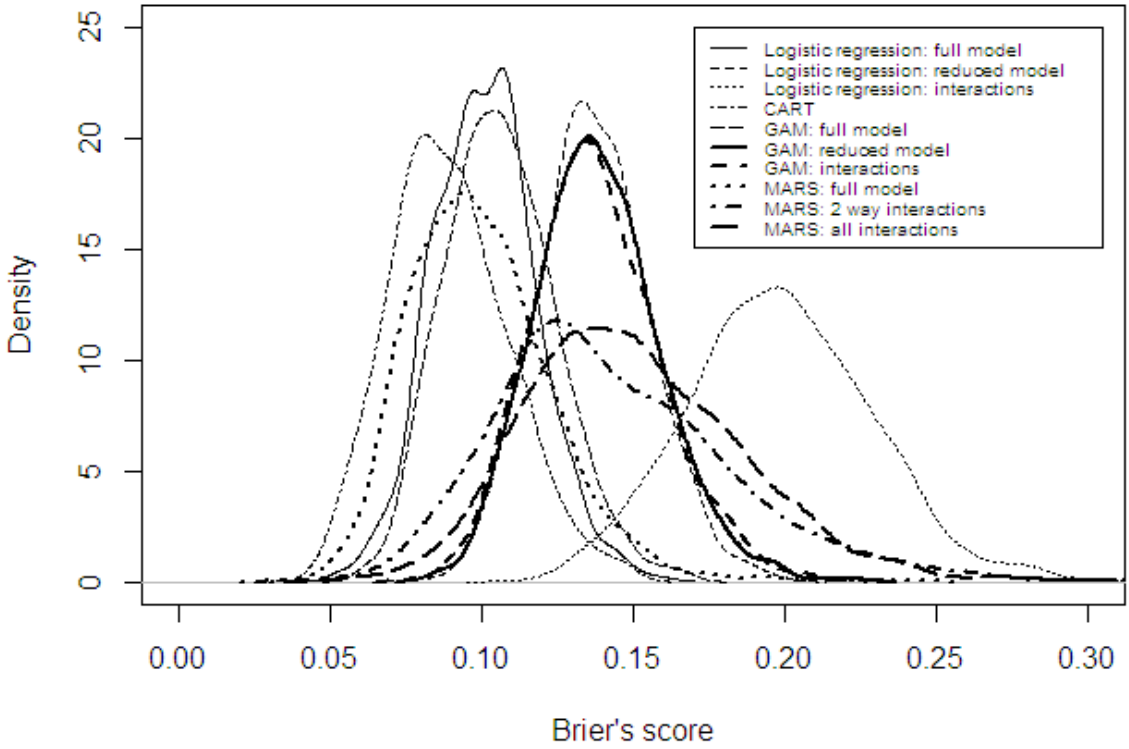

Figure 4: Distribution of sensitivity scores in 1000 validation samples

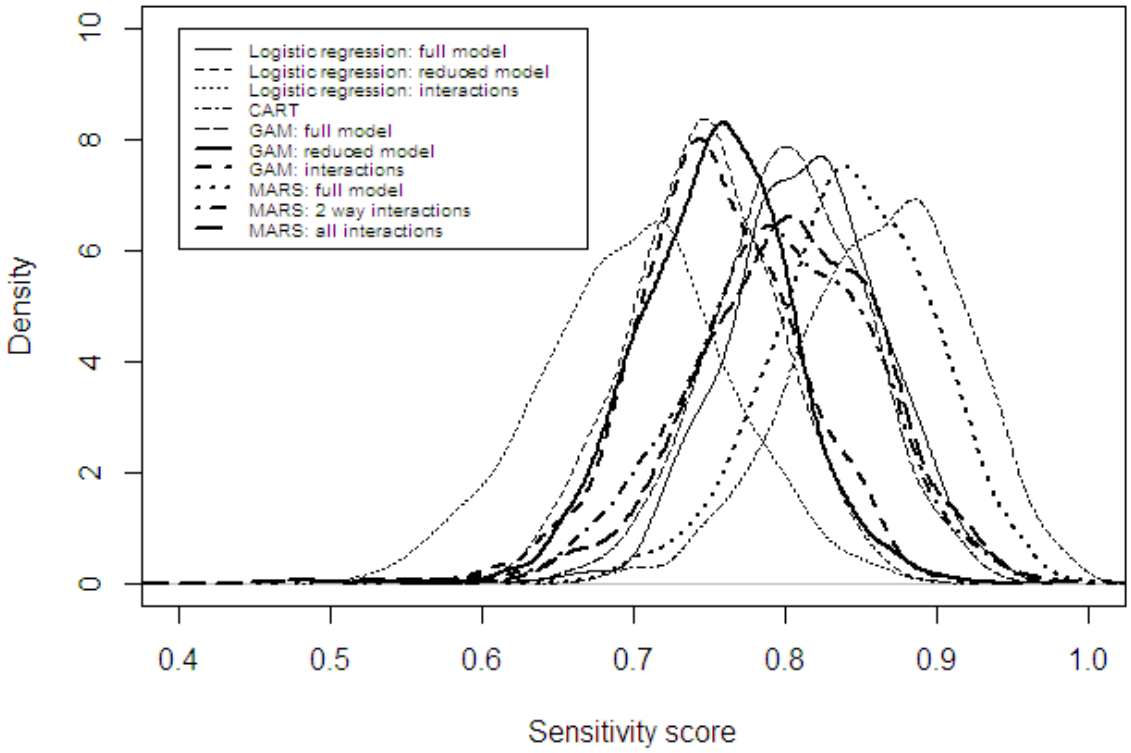

Figure 5: Distribution of specificity scores in 1000 validation samples

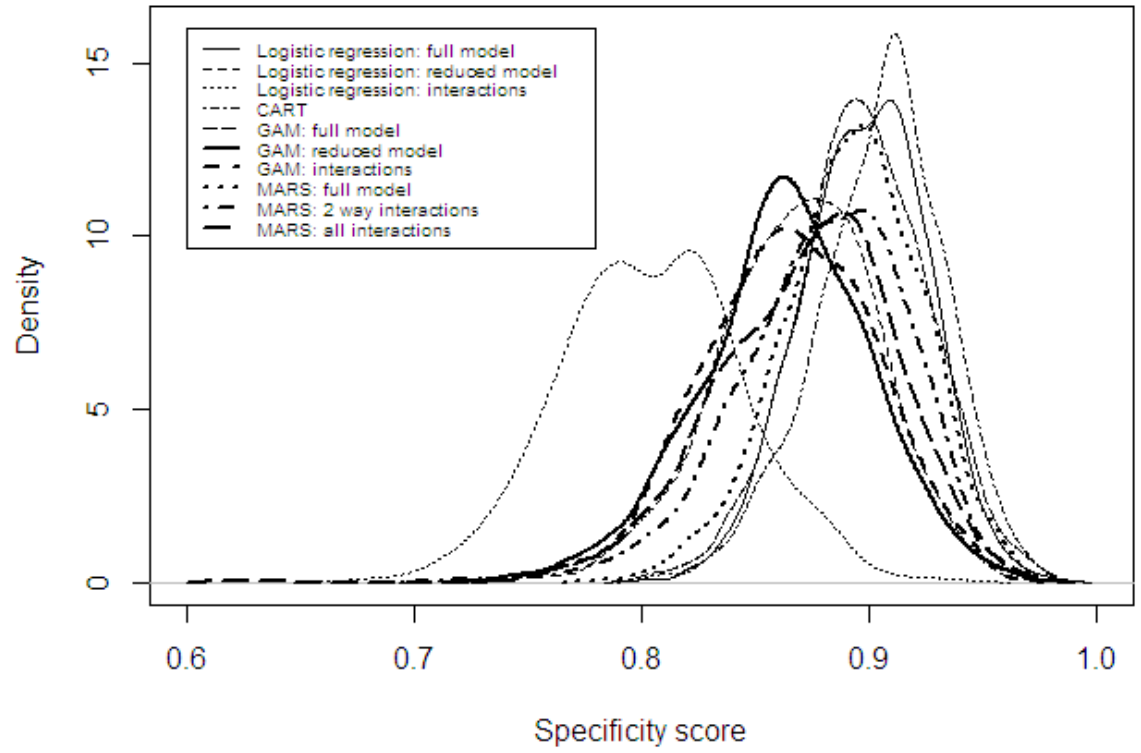

Figure 6: Regression tree for patients at risk of readmission 45 days after discharge. Each node contains the number of patients in that node and the risk of readmission rate of those patients [N (risk of readmissions)]. The derivation sample has 725 patients of which 45% were readmitted within 45 days after discharge.

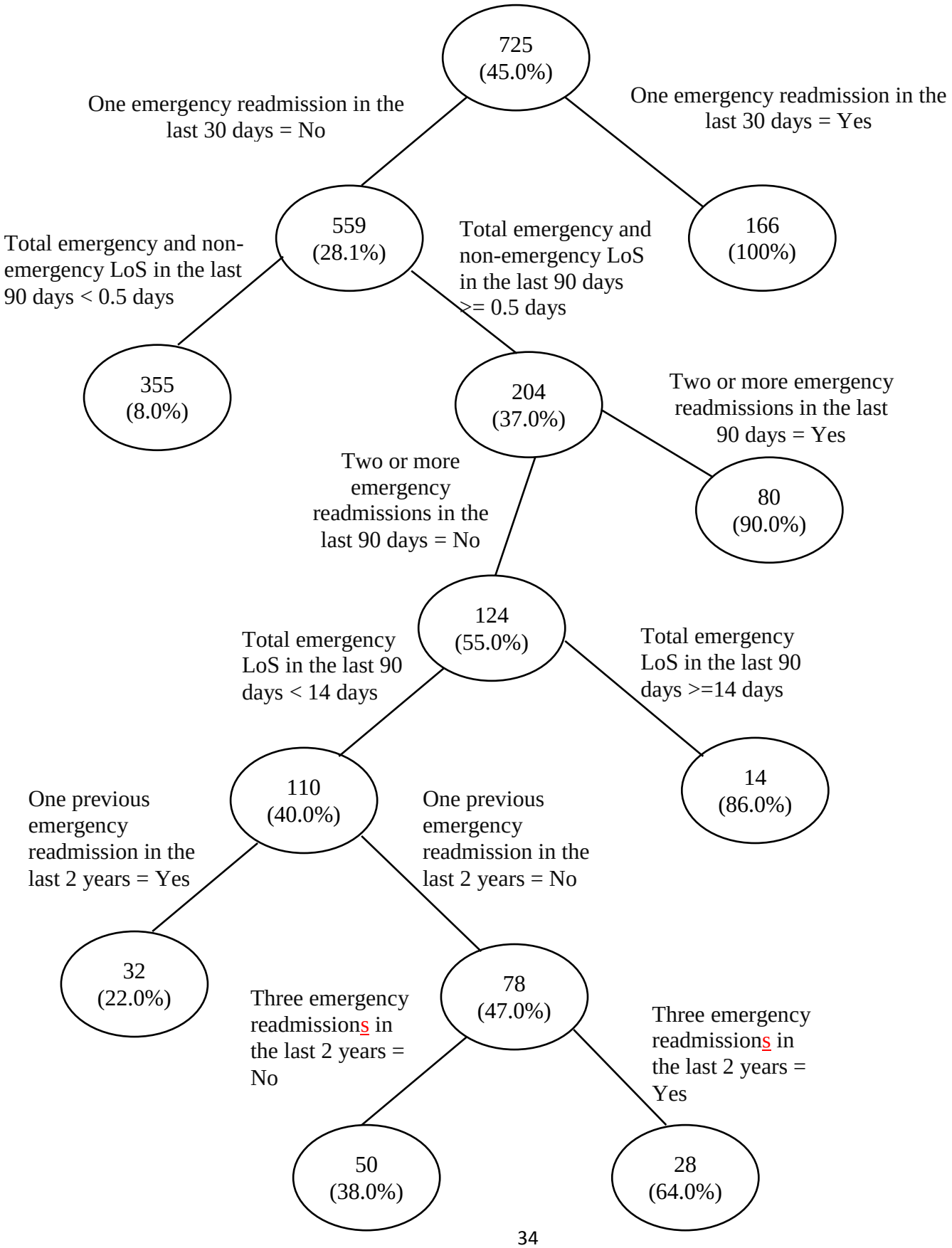

Figure 7: Relationship between selected variables and risk of readmission: generalized additive models (full model)

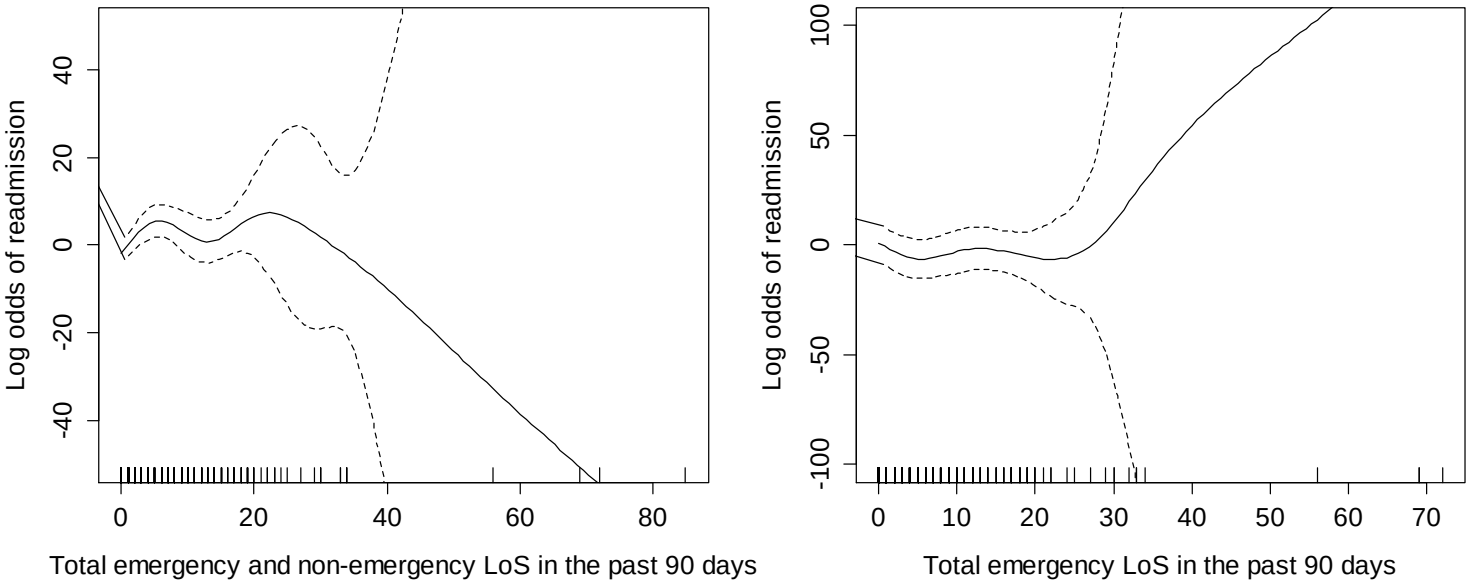

Figure 8: Relationship between selected variables and risk of readmission: multivariate adaptive regression splines (full model)