# 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, Business School, University of Hertfordshire, Hertfordshire, UK, e-mail: e.demir@herts.ac.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 45 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 and asthma. 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 1,000 times—the initial data set was divided into derivation and validation samples 1,000 times, and the predictive accuracy of each method was assessed each time. The mean ROC curve area for the regression tree models in the 1,000 derivation samples was .928, while the mean ROC curve area of a logistic regression model was .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 (healthcare managers, policy makers, etc.) for informed decision making in the management of diseases, which ultimately contributes to improved

[†]Corresponding author.

***Subject Areas: Risk of readmission, COPD, and Predictive Analytics.***

## INTRODUCTION

Since the 1980s, the United Kingdom (U.K.) health service has undergone major changes in its organization and delivery. The rising cost of care, changes in technology, pressures associated with demographic change (e.g., aging population), and different patterns of health-seeking behavior have forced the U.K. government to adapt new strategies that encompass whole health provisions, of which hospitals seek optimal output under strict economical constraints. Improving the management of high-cost patients, especially those with long-term conditions (LTCs), is increasingly seen as an important strategy for improving health outcomes and controlling healthcare expenditure and is a key element of the current U.K. National Health Service (NHS) policy (Department of Health, 2010a).

From the perspective of healthcare costs, unplanned hospital readmissions have widely been viewed as one of the most serious problems in health services. In the 5-year period between April 1, 2004, and March 31, 2010, 7% of patients discharged from a hospital in England were readmitted to hospital within 30 days (Clarke, Blunt, & Bardsley, 2012), with costs to the NHS estimated at £1.6 billion each year (Billings et al., 2012). Given the rising cost of care, Department of Health guidance for the NHS proposes that commissioners should not pay provider hospitals for emergency readmission within 30 days of an index elective (planned) admission (Department of Health, 2011b). Furthermore, the rate of readmissions also plays an important part in monitoring health system performance, as one of the new English public health "outcome indicators (Department of Health, 2010b).

Many different interventions have been introduced with the aim of reducing unplanned admission rates (Purdy, 2011); the evidence for their efficacy and cost-effectiveness is limited (Hansen, Young, Hinami, Leung, & Williams, 2011). One reason why interventions may be unsuccessful is if they are offered to patients who are at insufficiently high risk of readmission (Steventon, Bardsley, & Billings, 2011). A history of recent hospital admissions is not an accurate predictor of future admissions by itself (Roland, Dusheiko, Gravelle, & Parker, 2005), and it seems that clinicians are often unable to make reliable predictions about which patients will be readmitted (Allaudeen, Schnipper, Orav, Wachter, & Vidyarthi, 2011).

Therefore, given the importance of emergency readmissions in the United Kingdom and other countries, in relation to long-term savings, health system performances, patient satisfaction and safety, there is a genuine need to accurately identify emerging risk patients, so that healthcare organizations could 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 personalized 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 an increasing body of literature attempting to describe and validate hospital readmission risk prediction tools (Kansagara et al., 2011). These models have been categorized into three:

i. models relying on retrospective administrative data (Holman, Preen, Baynham, Finn, & Semmens, 2005; Bottle, Aylin, & Majeed, 2006; Krumholz et al., 2008a, 2008b, 2008c; Howell, Coory, Martin, & Duckett, 2009; Bardhan, Kirksey, Oh, & Zheng, 2011; Hammill et al., 2011; Billings et al., 2012),

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

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

Most of these models had poor predictive ability, where the area under the receiver operating characteristic (ROC) curve ranged from .55 to .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 .83. Here, ROC is defined to be the proportion of times the model correctly discriminates a pair of readmitted and nonreadmitted patients. The area under the curve of .50 indicates that the model performs no better than chance; .70 to .80 indicates modest or acceptable discriminative ability; and a value of greater than .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 (Stewart & Stamm, 1991; Hasford, Ansari, & Lehmann, 1993; Long, Griffith, Selker, & D'Agostino, 1993; Nelson, Bloch, Longstreth, & Shi, 1998; Sauerbrei, Madjar, & Prompeler, 1998; Tsien, Fraser, Long, & Kennedy, 1998; Avila, Segal, Wong, Boushey, & Fahy, 2000; El-Solh, Sikka, & Ramadan, 2001; Kuchibhatla & Fillenbaum, 2003; Lemon, Roy, Clark, Friedmann, & Rakowski, 2003; Schwarzer, Nagata, Mattern, Schmelzeisen, & Schumacher, 2003; Nishida et al., 2005; James, White, & Kraemer, 2005). 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 objective of this article is to validate the variables derived from prior medical services consumptions such as outpatient and accident & emergency (A&E) attendances and inpatient admissions, which have not been used extensively in prior models, and compare several different predictive models on readmission, and test if there is a consensus across different validation measurements of predictive performances. We compare the predictive ability of conventional logistic regression and regression tree methods with that of two other data-driven methods known as generalized additive models (GAMs) and multivariate adaptive regression splines (MARS) models. Predictive models tend to assume that continuous independent variables (such as age and length of stay [LoS]) have a linear effect on risk of readmission, which may not always be true. For instance, the risk of readmission for ages 50–64 may decrease and gradually increase for the 65+ population. In this context, assuming a linear effect may produce inaccurate and unstable results. GAM and MARS are versatile approaches and relax this assumption by fitting a nonlinear smooth arbitrary function (e.g., lowess and smoothing splines) on continuous variables.

Note that the purpose of this article is to predict patients who are known to the system, that is, it consists of all readmitted patients and does not predict patients who are not readmitted, simply because these are the cohort of patients that are costly, have an impact on key performance indicators including mortality, waiting lists, cancellation of planned care, etc. For this reason we have not included patients who are not known to the system.

Given that the existing literature on patient readmissions has produced poor prediction accuracies, there is a dire need for accurate and robust models with high predictive accuracies. This has serious practical implications, simply because incorrectly identifying patients at risk (or not at risk) of readmission can lead to poor use of resources (hence very costly) at a time when health services around the world including the NHS in England are struggling to provide care efficiently and effectively under severe financial constraints. Fulfilling these objectives could possibly enable us to develop reliable models for practitioners and managers for improving service delivery, achieve key targets (through reduction in emergency readmissions), and cost savings.

These objectives have never been investigated within the readmission phenomenon. The novelty of this article will primarily be based on these objectives and the new insights offered from key findings.

We used a large sample of patients hospitalized with chronic obstructive pulmonary disease (COPD) and asthma in a primary care trust (PCT) in England. COPD and asthma are known to be one of the leading causes of emergency readmission in the United Kingdom (Roland et al., 2005). For the first time ever, four well-established methodologies are rigorously evaluated and compared using newly derived variables (that have never been 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, 2011a), and the Government's initiative toward 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. Even without monetary incentives, knowledge of readmission risk could still be useful to clinicians for focusing their discharge planning efforts and postdischarge support on high-risk patients. This research will also support positive patient-centered 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 (healthcare managers, policy makers, etc.) for informed decision making in the management of diseases, which ultimately contributes to improved measures for hospital performance management.

## METHODS

### Data Sources

The data provided by a PCT in England comprised three key sources of data: inpatient care, outpatient care, and A&E. The inpatient care data set 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 data set contains information on outpatient appointments to NHS hospitals (day cases). It includes appointment dates, attendance types and nonattendances, waiting times, clinical and geographical data, patient details, socioeconomic factors, referral source, and outcome results. The A&E data set 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 Web site (Hospital Episode Statistics, 2012).

The data were 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 data set is 275,366 records, 122,446 inpatient care admissions, and 1,022,113 outpatient attendances. The first 2 years are used to develop the predictive models (derivation sample) and the third year (validation sample) to evaluate the observed versus predicted results.

The inpatient data capture all the consultant episodes of patients during their stay in hospital. During a hospital stay a patient might encounter several successive episodes (i.e., consultant treatments), collectively known as a spell. Hence, 122,446 episodes are made up of 96,448 spells or in other words, visits. Out of 96,448 visits 19,750 readmissions were observed (see Table 1 for details).

The inpatient data are the focus data set where all readmissions are captured. We use the outpatient and the A&E data sets (in addition to the inpatient data) to observe if there is a prior use of medical services, which are then used as explanatory variables. Out of over 100 key articles based on emergency readmissions (Demir, 2008), we found that prior use of medical services were highly associated with emergency readmission. No research has considered the footsteps of patients in other services, that is, outpatient and A&E. For example, if a patient had two or more A&E admissions in the last 90 days prior to next readmission, this may have an impact on next readmission. Similarly, a sick patient would normally visit his/her consultant in outpatient services on a regular basis. Hence, if a patient had 6–10 outpatient visits in the last 730 days (prior to next readmission), this may also have an impact on next readmission. What about if they had an X-ray taken or arrived by an ambulance in the last 90 days. We derived just over 60 explanatory variables from the inpatient, outpatient, and A&E data sets (variables listed in Table 2). Therefore, the outpatient and the A&E data sets were used to extract information about individual patient's footsteps in other services.

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 United Kingdom (Roland et al., 2005). Readmission time is the time (in days) from the date of discharge to next emergency admission. In the literature, the definition of readmission varies according to the purpose of the study, generally from 30 to 90 days (Moloney, Bennett, & Silke, 2004; Kemper, Kennedy, Dechert, & Saint, 2005), but some studies have used readmissions following certain surgeries, for shorter (14 days) (Reed, Pearlman, & Buchner, 1991) or longer time window (1 year) (Datar & Sood, 2011). The chosen time windows are generally subjective. Instead of relying on a subjective definition based on an arbitrarily chosen time window (e.g., 30 days), we objectively defined the time window for COPD and asthma as 45 days (Demir & Chaussalet, 2011). We capture the readmission process from the community using a special case of a Coxian phase-type distribution, expressed as a mixture of two generalized Erlang distributions. Using the Bayes theorem, we compute the optimal time windows in defining readmission. Therefore, readmission is defined to be all emergency admissions within 45 days after discharge from hospital for the same primary diagnosis (i.e., patient level analysis). Readmissions within 45 days were restricted according to the provisions of the 2011–2012 NHS operating framework by excluding nonemergency admissions

**Table 1:** A breakdown of the datasets and readmissions.

| Data | Total of Records | Total # of Reads (All Diseases) | Total # of Reads (COPD and Asthma Only) | Total # of Reads After Data Cleansing Process[a] | Derivation Sample[b] | Validation Sample[c] |
|---|---|---|---|---|---|---|
| Inpatient | 122,446 | 19,750 | 1110 | 963 [413, 550] | 725 [323,402] | 238 [90, 148] |
| Outpatient | 1,022,113 | N/A | N/A | N/A | N/A | N/A |
| A&E | 275,366 | N/A | N/A | N/A | N/A | N/A |

[a]413 and 550 are the total number of high- and low-risk group patients, respectively (readmitted within 45 days after discharge and greater than 45 days, respectively).
[b]The first 2 years is the derivation sample, that is, 725 readmissions (out of 963) of which 323 and 402 are the high- and low-risk group patients, respectively.
[c]The third year (or .75 year) is the validation sample, that is, 238 readmissions (out of 963) of which 90 and 148 are the high- and low-risk group patients, respectively.

**Table 2:** 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 LTC (0/1) | 54.1% |
| Two or more LTCs (0/1) | 14.8% |
| Two distinct inpatient primary diagnosis (0/1) | 22.6% |
| Three distinct inpatient primary diagnosis (0/1) | 10.4% |
| Four and above distinct in-patient primary diagnosis (0,1) | 22.2% |
| *Prior use of medical services: Inpatient care* | |
| One emergency admission in the past 30 days (0/1) | 22.7% |
| More than one emergency admission in the past 30 days (0/1) | 5.4% |
| One emergency admission in the past 90 days (0/1) | 32.5% |
| Two or more emergency admissions in the past 90 days (0/1) | 19.1% |
| One emergency admission in the past 180 days (0/1) | 33.2% |
| Two or more emergency admissions in the past 180 days (0/1) | 33.7% |
| One emergency admission in the past 365 days (0/1) | 35.1% |
| Two emergency admissions in the past 365 days (0/1) | 14.7% |
| Three or more emergency admissions in the past 365 days (0/1) | 30.9% |
| One emergency admission in the past 730 days (0/1) | 35.1% |
| Two emergency admissions in the past 730 days (0/1) | 15.8% |
| Three emergency admissions in the past 730 days (0/1) | 9.0% |
| Four or more emergency admissions in the past 730 days (0/1) | 28.2% |
| Total previous emergency length of stay (LoS) prior to emergency admission in the last 30 days | 0.9 (0–0) |
| Total previous emergency LoS prior to emergency admission in the last 90 days | 3.3 (0–3) |
| Total previous emergency LoS prior to emergency admission in the last 180 days | 6.0 (0–7) |
| Total previous emergency LoS prior to emergency admission in the last 365 days | 10.2 (0–12) |
| Total previous emergency LoS prior to emergency admission in the last 730 days | 14.6 (1–16) |
| Total previous emergency and nonemergency LoS prior to emergency admission in the last 30 days | 1.0 (0–0) |
| Total previous emergency and nonemergency LoS prior to emergency admission in the last 90 days | 3.5 (0–4) |
| Total previous emergency and nonemergency LoS prior to emergency admission in the last 180 days | 6.4 (0–8) |
| Total previous emergency and nonemergency LoS prior to emergency admission in the last 365 days | 10.8 (0–13) |
| Total previous emergency and nonemergency LoS 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% |
| Two previous emergency readmission as high-risk group in the last 730 days (0/1) | 24.8% |

*Continued*

**Table 2:** Continued

| *Medical Comorbidity* | |
|---|---|
| Three previous emergency readmission as high-risk group in the last 730 days (0/1) | 17.4% |
| Four previous emergency readmission as high-risk group in the last 730 days (0/1) | 12.7% |
| Five previous emergency readmission as high-risk group in the last 730 days (0/1) | 9.7% |
| ***Prior use of medical services: Outpatient care*** | |
| One outpatient specialty visit in the last 30 days (0/1) | 15.9% |
| Two outpatient specialty visit in the last 30 days (0/1) | 4.4% |
| Three or more outpatient specialty visit in the last 30 days (0/1) | 2.4% |
| One outpatient specialty visit in the last 90 days (0/1) | 20.8% |
| Two outpatient specialty visit in the last 90 days (0/1) | 11.4% |
| Three or more outpatient specialty visit in the last 90 days (0/1) | 13.9% |
| 1–5 outpatient specialty visits in the last 730 days (0/1) | 38.7% |
| 6–10 outpatient specialty visits in the last 730 days (0/1) | 19.7% |
| Eleven or more outpatient specialty visits in the last 730 days (0/1) | 15.2% |
| ***Prior use of medical services: Accident & Emergency (A&E)*** | |
| If the patient had an X-ray in their A&E visit in the last 180 days (0/1) | 53.0% |
| Arrived by ambulance in the last 90 days (0/1) | 41.2% |
| The patient was discharged to hospital in the last 180 days (0/1) | 53.1% |
| One A&E visit in the last 365 days (0/1) | 16.1% |
| Two A&E visit in the last 365 days (0/1) | 14.5% |
| Three or more A&E visit in the last 365 days (0/1) | 55.2% |
| ***Patient characteristics, sociodemographic and social determinants*** | |
| Age group 0–4 (0/1) | 6.0% |
| Age group 5–14 (0/1) | 4.5% |
| Age group 15–39 (0/1) | 9.3% |
| Age group 40–59 (0/1) | 18.5% |
| Age group 60–64 (0/1) | 6.7% |
| Age group 65–69 (0/1) | 8.4% |
| Age group 70–74 (0/1) | 13.2% |
| Age group 75–79 (0/1) | 10.5% |
| Age group 80–84 (0/1) | 14.1% |
| Age group 85–89 (0/1) | 6.5% |
| Age group 90–94 (0/1) | 2.2% |
| Age 95+ (0/1) | 0.1% |
| Gender (female) (0/1) | 51.2% |
| Age (continuous variable) | 60 (51–79) |
| IMD (continuous variable) | 24.9 (18–31) |
| ***Ethnicity*** | |
| British (White) (0/1) | 63.9% |
| Irish (White) (0/1) | 3.3% |
| Any other white background (0/1) | 1.3% |
| White and Black Caribbean (Mixed) (0/1) | 0.4% |
| White and Black African (Mixed) (0/1) | 0.3% |
| White and Asian (Mixed) (0/1) | 0.2% |

*Continued*

**Table 2:** Continued

| *Medical Comorbidity* | |
| --- | --- |
| Indian (Asian or Asian British) (0/1) | 14.1% |
| Pakistani (Asian or Asian British) (0/1) | 3.3% |
| Bangladeshi (Asian or Asian British) (0/1) | 0.1% |
| Any other Asian background (0/1) | 3.7% |
| Caribbean (Black or Black British) (0/1) | 0.3% |
| African (Black or Black British) (0/1) | 0.8% |
| Any other Black background (0/1) | 0.8% |
| Chinese (other ethnic group) (0/1) | 0.2% |

An LTC is defined to be COPD, asthma, coronary artery disease, congestive heart failure, hypertension, and cancer.

"One and only one LTC (0/1)" refers to patients who have a single LTC and no other LTCs; 54.1% of all patients have a single LTC. "Two or more LTCs (0/1)" refers to patients who have two or more LTCs (including COPD or asthma). For instance, if a patient had COPD and hypertension, then he/she is assigned a value of 1 for this particular variable (else 0).

A primary diagnosis is defined as the main condition treated. "Two distinct inpatient primary diagnosis (0/1)" refers to patients who had exactly two main conditions treated in the past. So, if the patient had two unique conditions treated in the past, then a value of 1 is assigned (else 0).

"Total previous emergency LoS prior to emergency admission in the last 180 days [6.0 (0–7)]" is the total number of days the patient stayed in hospital in the last 180 days (prior to next readmission) as an emergency admission; 6.0 is the average LoS in hospital and 0 and 7 (numbers in brackets) are lower and upper quartiles, respectively.

"Two previous emergency readmission as high-risk group in the last 730 days (0/1)" refers to patients who were readmitted twice as a high-risk group patient (i.e., readmitted within 45 days after discharge) in the last 730 days prior to next readmission. So, 24.8% of all patients were readmitted twice in the last 2 years.

and admissions where a national tariff was not applicable. 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. The 122,446 inpatient episodes account for 96,448 spells (or visits) with 19,750 patient readmissions (all clinical conditions), so almost 20% of visits to inpatient care were readmissions. COPD and asthma readmissions accounted for 1,110 readmissions (approximately 5.6% of all readmissions). After the 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 reduced to 963 (413 and 550 in the high risk and low risk group of readmission, respectively). Furthermore, the first 2 years of the data were used as the derivation sample (725) and the remaining for validation purposes (238). From the derivation sample there were 323 and 402 high- and low-risk groups of readmission, respectively, whereas 90 and 148, respectively, for the validation sample (see Table 1 for a breakdown of the data sets and readmissions).

All variables listed in Table 2 were derived through database programming. The variables are categorized into four areas: medical comorbidity, prior use of

medical services, patient characteristics, sociodemographic and social determinants. Around 14.8% of patients were diagnosed with two or more LTCs (Table 2). An 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 LoS prior to emergency admission in the last 30 days is around 6 days (25th and 75th percentile is 0 and 7 days, respectively). Interestingly, almost a fifth of patients had 6–10 outpatient attendances in the past 2 years. Furthermore, 55% had three or more A&E visits in the past 1 year.

## 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% of the bootstrap samples were retained for inclusion in the final predictive model. The resultant model comprised 38 variables (listed in Table 3), that is, 38 were found to be significant in at least 75% of the bootstrap samples out of the 79 variables listed in Table 2. This model will be used as the basis for some of the regression models that we will consider in this study.

## 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).

### *Logistic regression*

Three separate logistic regression models were developed to predict patients at risk of readmission. The first model consisted of the 38 variables described above, known as the reduced model. The second model was constructed using backward variable elimination. This consisted of all the 38 variables along with all two-way interactions, with the 38 main effects being forced to remain in each model. The third model was also constructed using backward variable elimination. However, in this instance, the initial model consisted of the 79 variables listed in Table 2. The logistic regression models were fitted using the "glm" function in R.

Backward variable elimination was done using the "step" function in R. This implementation of backward 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

**Table 3:** List of variables for the reduced model.

One and Only One Long-Term Condition (LTC) (0/1)

Two or more LTCs (0/1)
Two distinct inpatient primary diagnosis (0/1)
Three distinct inpatient primary diagnosis (0/1)
Four and above distinct inpatient 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 (LoS) prior to emergency admission in
    the last 90 days
Total previous emergency LoS prior to emergency admission in the last 180
    days
Total previous emergency LoS prior to emergency admission in the last 365
    days
Total previous emergency LoS prior to emergency admission in the last 730
    days
Total previous emergency and nonemergency LoS prior to emergency
    admission in the last 90 days
Total previous emergency and nonemergency LoS prior to emergency
    admission in the last 180 days
Total previous emergency and nonemergency LoS prior to emergency
    admission in the last 365 days
Total previous emergency and nonemergency LoS 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 outpatient specialty visit in the last 30 days (0/1)
1–5 outpatient specialty visits in the last 730 days (0/1)
11 or more outpatient 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 an LTC in the past (0/1)

prespecified boundary model is achieved or when no step will cause a further reduction in the AIC criterion (Hastie & Pregibon, 1993).

### *Classification trees*

Binary recursive partitioning methods are rarely used to construct regression trees to predict patients at risk of readmission. Classification tree–based models are conceptually simple yet powerful (Hastie, Tibshirani, & Friedman, 2009) and is known to be a data-intensive approach and generally works very well with large data sets (large in terms of the number of observations and the number of independent variables). 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 2. 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.

### *GAMs*

A GAM is an additive regression model of the form

$$y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \varepsilon, \tag{1}$$

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 because 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,

$$y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + Z_\gamma + \varepsilon, \tag{2}$$

where $Z$ is the design matrix for the variables that will not be modeled additively, where some may be quantitative and others qualitative. The $\gamma$'s are the associated regression parameters.

In this 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 3). Total LoS variables and age at admission were modeled 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 modeled using smoothing splines. The GAMs were fitted using the "gam" function in the MGCV package in R.

### *MARS models*

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 2. 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 overfitting the model and 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 overfit to the derivation sample. The MARS models were fit using the "earth" function in the EARTH package.

## 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 2 years of data were used for model derivation and the remaining 9 months of data were used for model validation. Each derivation sample consisted of 725 patients, while each validation sample consisted of 238 patients. This process was repeated 1,000 times. Chi-square tests and *t*-tests were carried out on a number of focal variables (medical comorbidity, prior use of medical services, and other characteristics) to examine whether the samples (derivation and validation) are equal. A total of 20 variables were selected of which two thirds were found to be equal (i.e., *p*-value > .05). 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

$$B = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{P}_i - Y_i \right)^2 , \qquad (3)$$

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 nonreadmitted cases. These two measures are well known by clinical and managerial staff in the NHS.

The above methods were repeated 1,000 times: the initial data were divided into derivation and validation components 1,000 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 1,000 validation data sets. By using 1,000 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.

It is highly recommended that predictive models should be evaluated using more than one performance metrics (ROC, sensitivity, specificity, Brier's Score, and generalized $R_N^2$ index) and see if there is an agreement between these measures. In addition, each of these measures have a unique interpretation, for instance, generalized $R_N^2$ index is very similar to the coefficient of determination (or adjusted $R^2$) in multiple linear regression, that is, the proportion of variability explained by the covariates, hence the generalized $R_N^2$ index has a direct statistical relevance. High ROC statistic, sensitivity, and specificity may have a direct relevance in predictive analytics (i.e., the proportion of cases correctly identified); however, it does not mean that the proportion of variability explained is also high, hence additional performance measures complement the traditional ones (ROC c statistics, sensitivity, specificity).

## RESULTS

### Predictive Performance

The mean area under the ROC curve for each model in both derivation and validation samples are reported in Table 4.

In the validation sample, the mean ROC curve area for the regression tree model was .924, while the mean ROC curve area for the stepwise logistic regression model was .928. The difference in ROC curve areas for the regression tree method

**Table 4:** Model discrimination in the 1,000 repeated split samples.

| Model | Receiver Operating Characteristic (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 | .948 | .924 | .721 | .089 | .862 | .904 |
| Logistic regression (backward elimination from full model) | .977 | .928 | .859 | .101 | .821 | .897 |
| Logistic regression (reduced model) | .954 | .880 | .759 | .137 | .773 | .868 |
| Logistic regression (two-way interactions) | .976 | .854 | .701 | .198 | .701 | .804 |
| GAM (full model) | .978 | .924 | .854 | .106 | .802 | .896 |
| GAM (reduced model) | .959 | .875 | .778 | .138 | .753 | .863 |
| GAM (two-way interactions) | .963 | .856 | .792 | .149 | .744 | .851 |
| MARS (full model) | .978 | .924 | .721 | .142 | .813 | .892 |
| MARS (two-way interactions) | .991 | .863 | .712 | .143 | .794 | .879 |
| MARS (all interactions) | .995 | .824 | .701 | .151 | .801 | .872 |

**Figure 1:** Distribution of receiver operating characteristic (ROC) curve areas in 1,000 validation samples.
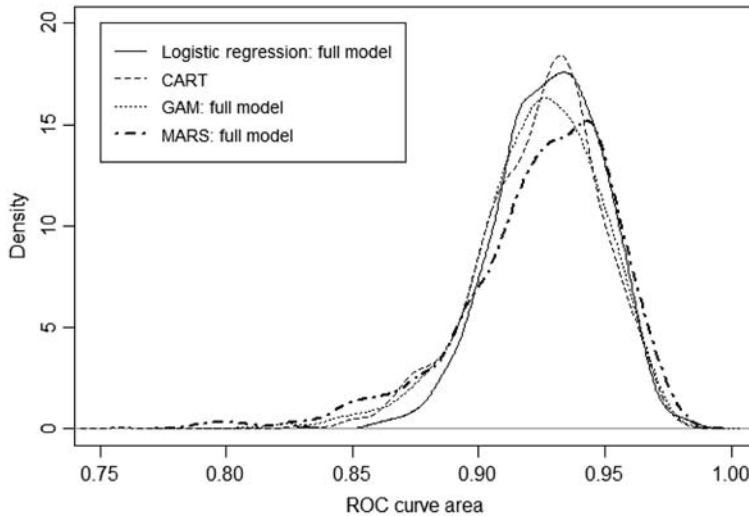


and the stepwise logistic regression model ranged from a low of zero to a high of .137 across the 1,000 validation samples (mean difference: .026). The ROC curve areas for the other modeling methods in the validation samples ranged from a low of .824 (MARS model with all interactions) to a high of .924 (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 overfitted on the derivation samples.

The mean ROC curve area for the regression tree decreased from .948 in the derivation samples to .924 in the validation samples—a very small decrease of .024. The decline in the mean ROC curve area between the derivation and validation samples was negligible for the logistic regression with the backward elimination from the full model (.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: .978 to .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 (.171), followed by MARS with two-way interactions (.131) and the logistic regression with two-way interactions (.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 overfit 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, the generalized $R_N^2$ index, Brier's score, sensitivity, and specificity score in the 1,000 validation data sets for promising models are described in Figures 1–3. Judging by the performance criteria from Table 4 the promising models are logistic regression (backwards elimination from full model), regression tree, GAM (full model), and MARS (full model). The

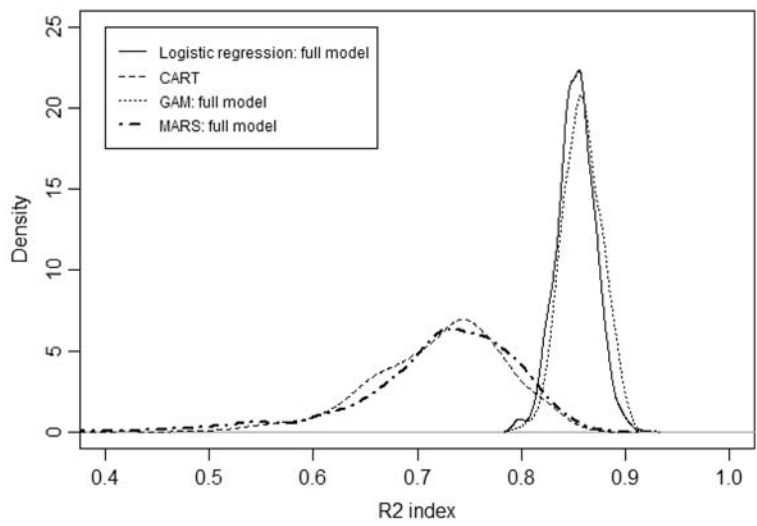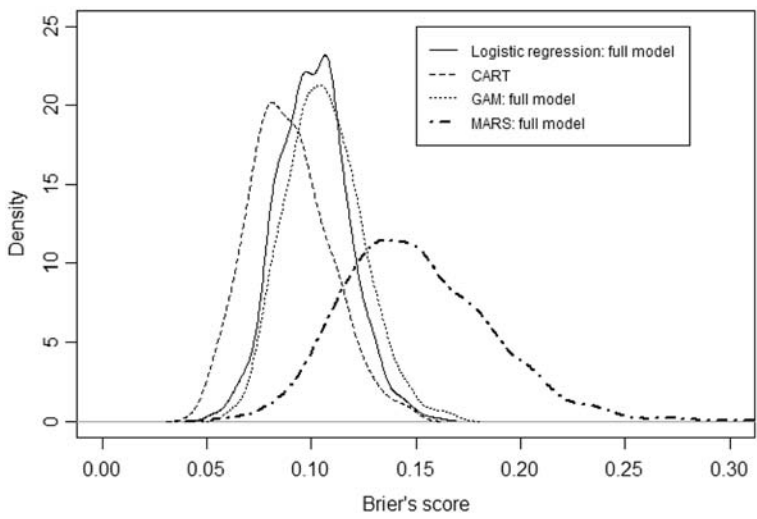**Figure 2:** Distribution of $R_N^2$ index in 1,000 validation samples.



**Figure 3:** Distribution of Brier's scores in 1,000 validation samples.



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 promising models (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. We notice that the distributions of the area under the ROC curve for promising models are very similar (Figure 1).

The generalized $R_N^2$ index is reported in Table 4 for each of the modeling strategies. The index ranged from a low of .701 for MARS with all interactions (and logistic regression with two-way interactions) to a high of .859 for the logistic regression model obtained from the full model using backward elimination, that is, this model explained substantial proportion of the observed variation. The Brier's score is also reported in Table 4 for each of the modeling strategies. Of note is the fact that there isn't too much variability in the estimated Brier's score for all the modeling 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 1,000 validation data sets for the promising models are described in Figures 2 and 3, respectively.

The distribution of the generalized $R_N^2$ index (Figure 2) in the validation samples shows that logistic regression (full model) and GAM (full model) have lower variability (and shifted to the right), which means that these two models have explained higher variability of the observed variation compared to the other two promising models (regression tree and MARS full model). In the case of the distribution of Brier's score (Figure 3), regression tree exhibited the least variability with the most effective predictive ability, whereas MARS (full model) shifted the most to the right (i.e., the least effective model) with high variability. The remaining two promising models can be considered to be comparable (.137 to .151).

Sensitivity and specificity scores are also reported in Table 4. The highest mean sensitivity and specificity scores over the 1,000 validation samples is from the regression tree model (.862 and .904, respectively) and logistic regression based on backward elimination from the full model (.821 and .897, respectively).

## Miscellaneous Results

The best modeling strategy is selected from each method and key findings are presented here. Judging from the predictive performance measures from Table 4, the stepwise logistic regression, GAM (full model), and MARS (full model) are selected. The results from the analysis have four important cases of interest for key decision makers (clinicians, managers). The number of correctly identified readmitted cases; the number of cases observed to be readmitted but predicted to be not readmitted; the number of correctly identified not-readmitted cases; and the number of cases observed to be not readmitted but predicted to be readmitted. Table 5 shows these figures for each of the selected models, where the figures are based on the validation sample (238 cases in total of which 90 were readmitted and 148 were nonreadmitted). Regarding the purpose of this research (i.e., reducing readmissions through tailored cares for patients at a high risk), correct predictions for readmission, and predicted nonreadmitted but observed to be readmitted are important.

Considering effective budgeting, a predicted readmission, but no readmission is a critical factor, because it makes a hospital pay additional costs, which are not actually necessary. Note that the predictions in Table 5 are very similar where they slightly differ in relation to other performance measures (see Table 4).

The regression tree obtained using one of the derivation samples is illustrated in Figure 4, which has seven terminal nodes. Six variables were used in creating

**Table 5:** The number of correctly identified readmitted cases; observed to be readmitted but predicted to be nonreadmitted; correctly identified nonreadmitted cases, and the number of cases observed to be nonreadmitted but predicted as readmitted. These figures are based on the validation sample (238 cases in total of which 90 were readmitted and 148 were nonreadmitted).

| Predicted Observed | Stepwise Logistic Regression | | | Regression Tree | | |
|---|---|---|---|---|---|---|
| | *Readmitted* | *Nonreadmitted* | *Sensitivity/ Specificity* | *Readmitted* | *Nonreadmitted* | *Sensitivity/ Specificity* |
| Readmitted | 74 | 16 | .821 | 78 | 12 | .862 |
| Nonreadmitted | 15 | 133 | .897 | 14 | 134 | .904 |
| | GAM (full model) | | | MARS (full model) | | |
| | *Readmitted* | *Nonreadmitted* | *Sensitivity/ Specificity* | *Readmitted* | *Nonreadmitted* | *Sensitivity/ Specificity* |
| Readmitted | 72 | 18 | .802 | 73 | 17 | .813 |
| Nonreadmitted | 15 | 133 | .896 | 16 | 132 | .892 |

**Figure 4:** 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.
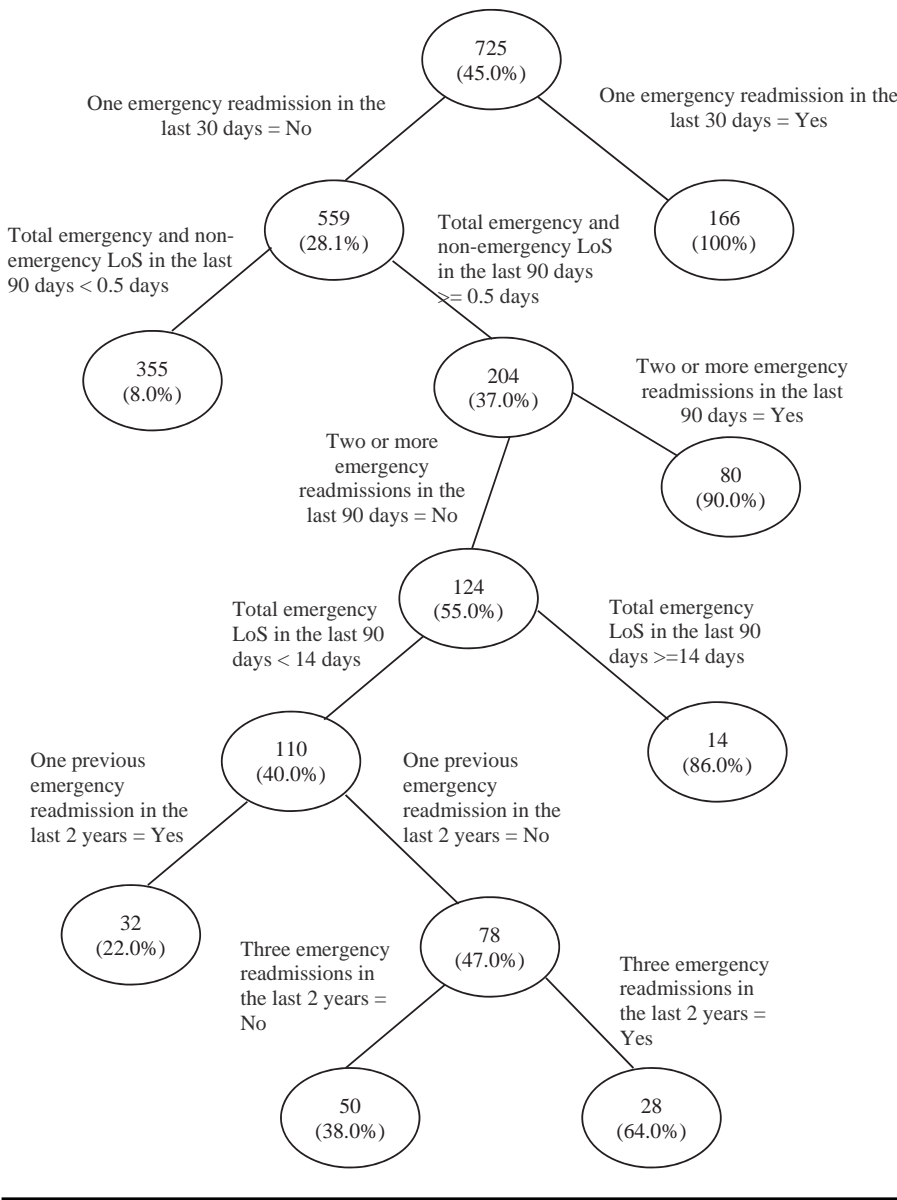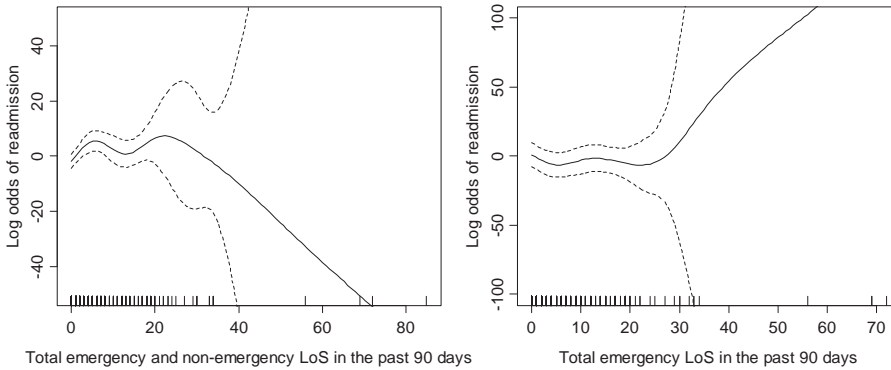
**Figure 5:** Relationship between selected variables and risk of readmission: generalized additive models (full model).



the tree, which are all related to prior use of medical services (e.g., LoS, previous history of readmissions). If a patient had experienced one previous emergency readmission (as a high-risk group patient) in the last 30 days then there is a 100% chance that this patient will be readmitted again within the next 45 days. Here, previous readmission is a prior use of medical services (the explanatory variable), whereas next readmission is part of the response variable. Note that the 30 days is not inclusive in the 45-day time window. 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 LoS in hospital (emergency and nonemergency) 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 backward elimination, 23 variables (out of a total of 79) were included in the final model, of which 19 had a *p*-value less than .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 5 describes the relationship between log odds of readmission and "total emergency and nonemergency LoS in the past 90 days" and "total emergency LoS 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 LoSs based on emergency admissions only, whereas the former includes nonemergency 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.

One observes that the relationship between the two variables and log-odds of readmission is nonlinear up to 30 days of LoS 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, the cumulative LoS of

**Figure 6:** Relationship between selected variables and risk of readmission: multivariate adaptive regression splines (full model).
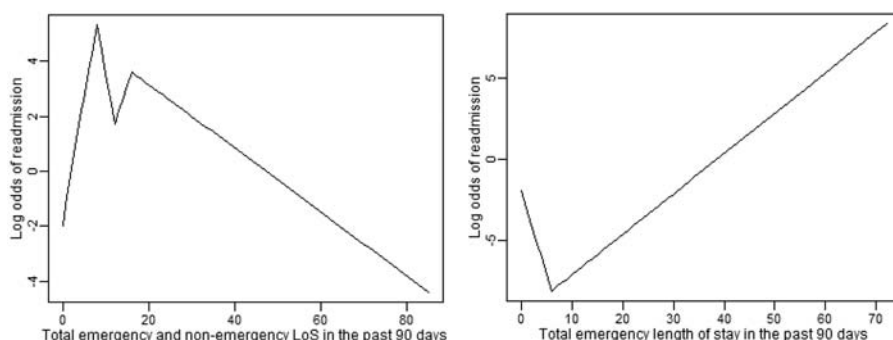


patients who were previously admitted as an emergency and nonemergency admission reduces the risk of being readmitted. One explanation for this phenomenon is that these patients are cared by clinicians and nurses on a regular basis (as an emergency and elective admissions via appointments) and when a patient is cared and treated for a prolonged period of time, our findings may suggest that increasing LoS in hospital lowers future risk of readmission.

Figure 6 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 nonemergency LoS) log odds of readmission increase rapidly, where this risk gradually decreases after an 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 6 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 United Kingdom. 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. Although there is no evidence to indicate that higher LoS reduces patient readmissions (or vice versa), Figures 5 and 6 gives us some insights to the issues surrounding LoS, that is, when a patient is cared and treated for longer periods of time (as an emergency and nonemergency LoS) the risk of future readmission may decrease. This finding should not be considered to be conclusive and needs to be further investigated.

## DISCUSSION

In the NHS, 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 (2010/2011), which represent 36.7% of hospital admissions in England (5.3 million admissions

in 2010/2011). Given that the tough economic conditions are expected to be with us for quite a while in the future, the U.K. Government's target is to provide personalized care plan for vulnerable people most at risk. Managing emergency readmissions will inevitably reduce the burden on nonemergency healthcare and resource use, which may lead to substantial amount of cost savings, reduction on waiting lists, and more importantly positive patient-centered 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 United Kingdom and other countries where these studies produced conflicting findings, resulting in poor predictions.

In this study, we have demonstrated, using a large sample of patients with COPD and asthma 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 were .928 and .924, respectively, in the validation sample, while the corresponding values for GAMs and MARS were .854 and .721, respectively. The highest observed ROC curve area was for the logistic regression model obtained from the full model using backward 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.

It is rather difficult to provide a guideline as to which performance measure to rely on when choosing a predictive methodology. We expect the optimal model to have an agreement between the performance measures, for instance, if the model performance has an ROC $c$-statistic greater than .85 (similar scores for sensitivity, specificity, and generalized $R^2$ index) and a Brier's score less than .15, then this model would be considered to be a reliable, robust, and an effective model (Harrell Jr, 2001; Austin, 2007). However, in the case that all model performance measures are similar then the simplest method should be selected for practical purposes, for example, logistic regression or the classification tree.

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 LoSs in the past 90 days" was nonlinear up to an 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 and asthma patients could pay particular attention to those patients who have been in emergency care for a total of 30 or more days. Note that 30 or more days refer to the cumulative LoSs 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 is 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.

A limitation of this study is that we have only considered those patients who are known to the system, that is, it consists of all readmitted patients and does not predict patients who are not readmitted. This may therefore restrict the generalization of the results because it is limited to patients with already highly significant medical resource consumptions. Future work will be directed toward the inclusion of the nonreadmitted cohort of patients.

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 nonreadmitted cases, respectively), as this approach has been criticized 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 1,000 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 45 days after discharge.

From the review of the literature, some of the risk factors examined were dependence, age, severity of illness, LoS, number of previous hospital admissions, mobility status, and care after discharge (Demir, 2008). Although there were many variables considered, no single factor could be used as a predictor of risk of readmission, except the number of prior hospital admissions was found to be a significant predictor (23 out of the 25 reviewed articles). Motivated by this finding, we tracked individual patient's footsteps in other services (outpatient, A&E, and inpatient) and derived many new variables (listed in Table 2) based on prior use of medical services. Judging by the results we found that these variables were the only predictors of risk of readmission and that none of the patient characteristic variables are listed as significant predictors. Furthermore, we found similar set of variables as predictors of readmission for other long-term care conditions, such as cancer, congestive heart failure, stroke, coronary artery disease, and hypertension, and thus our results are conclusive and in line with the literature, except with new variables offered for researchers and practitioners to utilize for risk prediction purposes. The inclusion of the new variables increased the performance dramatically. For instance, we tested the existing predictive tool (developed by the King's Fund and

Nuffield Trust, Billings et al., 2006) with the same set of variables excluding the new ones and produced an ROC statistics of .65, whereas based on our model (including all new variables) the ROC statistics was greater than .9, hence a massive difference. Furthermore, the results were very promising for other disease categories as well including cancer, congestive heart failure, stroke, coronary artery disease, and hypertension.

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 to predict the risk of readmission (within 45 days after discharge) for patients that are already known to the system. 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.

## CONCLUSION

One of the most difficult and challenging aspects of predictive modeling is being able to identify the key sets of variables that are likely to have an impact on the prediction of readmissions and in most cases relevant data may not be available. A number of authors have investigated the impact of using administrative data (i.e., routinely collected) versus clinical data in profiling coronary artery by-pass grafting mortalities in New York hospitals, and found that models based upon clinical data provided better discrimination and accuracy in explaining variations in patient mortality (Hannan, Kilburn, Lindsey, & Lewis, 1992). Unfortunately the data found in almost all national healthcare databases tend to be highly administrative, and given that the data provided to us by a CCG are administrative, substantial amount of time was dedicated toward deriving new set of variables (over 100) through database programming, where each variable was tested for its significance and the impact it had on the predictive power of models. Therefore, without clinical data, the newly derived variables have made a significant improvement on model performance and above all contributed toward the development of robust and reliable predictive models.

The existing literature highlights the lack of models that are able to correctly identify patients at risk of readmission. In this respect, incorrect identification of patients at risk (or not at risk) of readmission can lead to poor use of resources (hence very costly) at a time when health services around the world including the NHS in England are struggling to provide care efficiently and effectively under severe financial constraints. This article will therefore inevitably enable many researchers around the world to use these new variables in their predictive modeling exercises (and the development of robust decision support tools). Furthermore, these variables are not just for predicting patient readmissions but could also be

utilized for other adverse outcomes, such as mortality, risk of heart failure, disease progression.

A number of CCGs in England (payers of care) have started to incorporate the predictive algorithms shown in this article as part of their analytical toolkit to ensure the following: (i) readmission is used as a performance indicator in England, so the lower the rate of readmission, the better for the provider (i.e., hospital), hence implementing the optimal model within a decision support systems framework will certainly enable hospitals to achieve key targets; (ii) cost savings, that is, the more patients are treated for care in the community, the less likely these patients will be readmitted back to hospital; and (iii) reduction in waiting lists, that is, release beds for elective admissions. Furthermore, we have developed similar models for cancer, coronary artery disease, stroke, congestive heart failure, and hypertension (the main chronic diseases), and the results are very similar. All the results have been rigorously evaluated, providing robust and accurate sets of findings.

Many published papers have compared logistic regression with classification trees, neural networks, and support vector machines; however, no attempt has been made to compare the traditional approaches with more sophisticated data-driven approaches (GAM and MARS). Predictive models tend to assume that LoS has a linear effect on log-odds of readmission, which may not always be true, as it is in our case here. GAM and MARS are versatile approaches and relax this assumption by fitting a nonlinear smooth arbitrary function (e.g., lowess and smoothing splines) on LoS. This is more flexible than the logistic regression model and is still interpretable because the functions $f_j$ can be plotted to give a sense of the marginal relationship between the predictors and the response (Figures 5 and 6). These approaches have provided some interesting insights, which can be beneficial for senior decision makers (clinicians, commissioners), for example, the issues surrounding emergency and nonemergency LoS.

Among all the models developed so far there have also been no attempts to assess the predictive power of these models. Using the repeated bootstrap re-sampling approach, we generated 1,000 validation samples and investigated the distribution of key performance measures, and clearly showcased the sensitivity of these measures based on many validation samples. Furthermore, the majority of studies have only used ROC statistics, sensitivity, and specificity scores, whereas in this article, we have considered other well-known performance measures (generalized $R_N^2$ index, Brier's score) and examined if there is a consensus among the measures to select the most appropriate model.

For these reasons the article should be considered a major contribution to the literature filling in a major gap in the literature.

## REFERENCES

Allaudeen, N., Schnipper, J. L., Orav, E. J., Wachter, R. M., & Vidyarthi, A. R. (2011). Inability of providers to predict unplanned readmissions. *Journal of General Internal Medicine*, 26, 771–776.

Amarasingham, R., Moore, B. J., Tabak, Y. P., Drazner, M. H., Clark, C. A., Zhang, S. et al. (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.

Bardhan, I., Kirksey, K., Oh, J., & Zheng, E. (2011). A predictive model for readmission of patients with congestive heart failure: A multi-hospital perspective. *Proceedings of the Thirty-second International Conference on Information Systems, Shanghai*, December 2011, 1–39.

Billings, J., & Mijanovich, T. (2007). Improving the management of care for high-cost Medicaid patients. *Health Affairs (Millwood)*, *26*(6), 1643–1654.

Billings, J., Blunt, I., Steventon, A., Georghiou, T., Lewis, G., & Bardsley, M. (2012). Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30). *British Medical Journal*, *2*, doi:10.1136/bmjopen-2012-001667.

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), 1–6.

Blunt, I., Bardsley, M., Grove, A.L. and Clarke, A. (2014). Classifying emergency 30 day readmissions in England using routine hospital data 2004–2010: what is the scope for reduction? Emergency Medicine Journal, doi:10.1136/emermed-2013-202531.

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.

Datar, A., & Sood, N. (2011). Impact of postpartum hospital-stay legislation on new-born length of stay, readmissions, and mortality in California. *Pediatrics*, *118*, 63–72.

Demir, E. (2008). Modelling Readmission: Defining and Developing a Framework for Profiling Hospitals. Ph.D Thesis. University of Westminster. U.K.

Demir, E., & Chaussalet, T. J. (2011). Capturing the re-admission process: A focus on time window. *Journal of Applied Statistics*, *38*, 951–961.

Department of Health. (2005). Case management competences framework for the care of people with long term conditions, accessed August 23, 2012,

available at http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4118101

Department of Health. (2010a). White paper: Equity and excellence-liberating the NHS, accessed August 1, 2012, available at http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/@ps/documents/digitalasset/dh_117794.pdf.

Department of Health. (2010b). *Healthy lives, healthy people: Transparency in outcomes, proposals for a public health outcomes framework*. London.

Department of Health. (2011a). Health and social care bill, accessed August 23, 2012, available at http://www.publications.parliament.uk/pa/cm201011/cmbills/132/11132.pdf.

Department of Health. (2011b). Payment by results guidance for 2011–12. London.

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*. Boca Raton: 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. et al. (2011). Incremental value of clinical data beyond claims to predict 30-day outcomes following heart failure admission. *Cardiovascular Outcomes Research*, *4*(1), 60–67.

Hannan, E. L., Kilburn, H. J., Lindsey, M. L., & Lewis, R. (1992). Clinical versus administrative databases for CABG surgery. Does it matter? *Medical Care*, *30*, 892–907.

Hansen, L. O., Young, R. S., Hinami, K., Leung, A., & Williams, M. V. (2011). Interventions to reduce 30-day rehospitalization: A systematic review. *Annals of Internal Medicine*, *155*, 520–528.

Harrell Jr, F. E. (2001). *Regression modeling strategies*. New York, NY: Springer.

Hasan, O., Meltzer, D. O., Shaykevich, S. A., Bell, C. M., Kaboli, P. J., Auerbach, A. D. et al. (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, NY: Chapman & Hall.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. London: Chapman & Hall.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd ed.). New York, NY: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. H., (2001). The Elements of Statistical Learning: Data Mining, Inference and Prediction (2nd Edition). Springer Series in Statistics. New York.

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, accessed August 24, 2012, available at 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. et al. (2011). Risk prediction models for hospital readmission. *Journal of the American Medical Association*, *306*(15), 1688–1698.

Kemper, A., Kennedy, E., Dechert, R., & Saint, S. (2005). Hospital readmissions for bronchitis. *Clinical Pediatrics*, *44*, 509–513.

Krumholz, H. M., Normand, S.-L. T., Keenan, P. S., Desai, M. M., Lin, Z., Drye, E. E. et al. (2008a). Hospital 30-day acute myocardial infarction readmission measure: Methodology, accessed August 20, 2012, available at 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. et al. (2008b). Hospital 30-day acute myocardial infarction readmission measure: Methodology, accessed August 20, 2012, available at 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. et al. (2008c). Hospital 30-day pneumonia readmission measure, accessed August 20, 2012, available at 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.

Moloney, E., Bennett, K., & Silke, B. (2004). Patient and disease profile of emergency medical readmissions to an Irish teaching hospital. *Postgraduate Medical Journal*, *80*, 470–474.

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. et al. (2005). Determination of smoking and obesity as periodontitis risks using the classification and regression tree method. *Journal of Periodontology*, *76*, 923–928.

Purdy, S. (2011). Avoiding hospital admissions: What does the research evidence say?, accessed July 12, 2012, Available at King's Fund: http://www.kingsfund.org.uk/publications/avoiding_hospital.html

R Core Development Team. (2005). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Reed, R., Pearlman, R., & Buchner, D. (1991). Risk factors for early unplanned hospital readmission in the elderly. *Journal of General Internal Medicine*, *6*, 223–228.

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.

Steventon, A., Bardsley, M., & Billings, J. (2011). *An evaluation of the impact of community-based interventions on hospital use*. London: Nuffield Trust. [Available at http://www.nuffieldtrust.org.uk/sites/files/nuffield/an-evaluation-of-the-impact-of-community-based-interventions-onhospital-usefull_report.pdf]

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. et al. (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.

**Eren Demir** received a BSc in Computational Mathematics, MSc in Decision Sciences, and a PhD in Applied Statistics in health and social care services. Since 2004 he has been lecturing in undergraduate and postgraduate statistics and operations research. He has published widely in the area of healthcare modeling and decision support systems. He is involved in the use of modeling techniques in healthcare management, and the development of tools that can aid healthcare decision making. He has participated in many consultancy/research projects, including capacity planning; measuring and modeling occupancy time in NHS continuing healthcare; forecasting length of stay and cost of long-term care for elderly people; redesign of services to improve clinical outcomes; data mining to capture hidden patterns and trends; simulation of patient pathways; evaluation of service outcomes, and development of models to measure cost effectiveness. He is currently a Senior Lecturer in statistics and operational research at the University of Hertfordshire and joint head of Healthcare Management and Policy Research Unit.