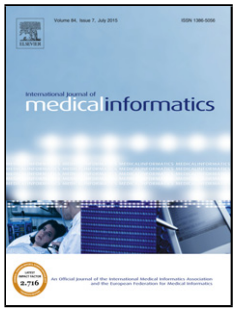
Ensemble Risk Model of Emergency Readmissions (ERMEA)

Mesgarpour, M., Chaussalet, T.J. and Chahed, S.



Title: Ensemble Risk Model of Emergency Admissions (ERMER)

Author: Mohsen Mesgarpour Thierry Chaussalet Salma Chahed

PII: S1386-5056(17)30088-6
DOI: http://dx.doi.org/doi:10.1016/j.ijmedinf.2017.04.010
Reference: IJB 3501

To appear in: *International Journal of Medical Informatics*

Received date: 21-12-2016
Revised date: 11-4-2017
Accepted date: 14-4-2017

Please cite this article as: Mohsen Mesgarpour, Thierry Chaussalet, Salma Chahed, Ensemble Risk Model of Emergency Admissions (ERMER), *<![CDATA[International Journal of Medical Informatics]]>* (2017), http://dx.doi.org/10.1016/j.ijmedinf.2017.04.010

Ensemble Risk Model of Emergency Admissions (ERMER)

Mohsen Mesgarpour[1,*], Thierry Chaussalet[1,*], Salma Chahed[1,*]

HSCMG, Faculty of Science and Technology, University of Westminster, 115 New Cavendish Street, W1W 6UW London, UK

Abstract

Introduction: About half of hospital readmissions can be avoided with preventive interventions. Developing decision support tools for identification of patients' emergency readmission risk is an important area of research. Because, it remains unclear how to design features and develop predictive models that can adjust continuously to a fast-changing healthcare system and population characteristics. The objective of this study was to develop a generic ensemble Bayesian risk model of emergency readmission.

Methods: We produced a decision support tool that predicts risk of emergency readmission using England's Hospital Episode Statistics inpatient database. Firstly, we used a framework to develop an optimal set of features. Then, a combination of Bayes Point Machine (BPM) models for different cohorts was considered to create an optimised ensemble model, which is stronger than the individual generative and non-linear classifications. The developed Ensemble Risk Model of Emergency Admissions (ERMER) was trained and tested using three time-frames: 1999-2004, 2000-05 and 2004-09, each of which includes about 20% of patients in England during the trigger year.

Results: Comparisons are made for different time-frames, sub-populations, risk cut-offs, risk bands and top risk segments. The precision was 71.6% to 73.9%, the specificity was 88.3% to 91.7% and the sensitivity was 42.1% to 49.2% across different time-frames. Moreover, the Area Under the Curve was 75.9% to 77.1%.

Conclusions: The decision support tool performed considerably better than the previous modelling

*Corresponding authors
Email addresses: mohsen.mesgarpour@gmail.com (Mohsen Mesgarpour), T.Chaussalet@westminster.ac.uk (Thierry Chaussalet), S.Chahed@westminster.ac.uk (Salma Chahed)

approaches, and it was robust and stable with high precision. Moreover, the framework and the Bayesian model allow the model to continuously adjust it to new significant features, different population characteristics and changes in the system.

Keywords: Hospital Episode Statistics, Readmission, Ensemble, Bayesian, Framework, Inpatient

1. Introduction

The cost of care is increasing at a rate that is unaffordable in the current economy. This is mainly due to the impact of ageing population, population growth, deprivations, the increase in emergency admissions, increased expectations, and the cost of treatment and technology (NHS, 2013; DH, 2013; Lewis et al., 2011). The current system is unsustainable and unfair, and the current financial options available to support people in meeting care costs are limited.

The National Health Service (NHS) spends an estimated £11 billion per year on emergency admissions in England (Lewis et al., 2011). According to the Nuffield Trust report in 2012 (Nuffield Trust, 2012), about 8% of discharged patients are readmitted within 30 days, costing an estimated £2.2 billion a year. Based on a retrospective study by Clarke et al. (2012) (Clarke et al., 2012), about half of the 30-day emergency readmissions were potentially preventable between 2004 and 2010.

Four major risks have contributed to the increase in emergency (or unplanned) readmissions to hospitals (HSCIC, 2013; Lewis et al., 2011): ageing population (Caley and Sidhu, 2011), patients with long-term conditions (DH, 2012), premature discharge and unpredictable accidents and emergency (Clarke et al., 2012). While discharging patients provides a way of freeing beds in healthcare systems, premature discharge could still increase the risk of emergency readmissions. Often hospital admission or readmission can be avoided by providing adequate care (Bardsley et al., 2012).

Therefore, developing and implementing a robust decision support tool for admitted patients is critical. Predictive risk models can help patients and carers obtain appropriate support services in clinical decision-making. In addition, such models can improve care quality and reduce the costs of inappropriate admissions to hospital and accident and emergency (A&E).

In 2005, the UK Department of Health (DoH) commissioned the Patients at Risk of Re-hospitalisation (PARR) (Lewis, 2011; Billings et al., 2006) algorithm and the PARR++ software for Primary Care

Trusts (PCTs) (Lewis et al., 2011; The King's Fund, 2016). The aim of the PARR model was to identify individuals at high risk of emergency readmission to a hospital within a year based on the inpatient data from the Hospital Episode Statistics (HES) database. Thereafter, in 2006, to address the need for identifying the patient risk along a continuum, the DoH released the Combined Predictive Model (CPM) which was based on General Practice (GP) and the HES data (DH, 2006).

In 2011, the DoH commissioned an upgrade to the PARR and the CPM models (Nuffield Trust, 2012; DH, 2011). The Patients at Risk of Readmission within 30 days (PARR-30) model was developed as an upgrade to be run by acute hospitals. The PARR-30 model was based on a broad range of measures used in the PARR (Billings et al., 2012).

After the controversies of the 2012 Heath and Social Care Act (Timmins, 2013), the care system moved towards developing new models of integrated care. The NHS's strategic five-year forward view (NHS, 2014) outlines that commissioners, the NHS and other providers will co-design the services based on a model of integrated care that targets specific cohorts, with their own exemplars, potential benefits, risks and transition cost.

In the NHS, patients' interactions with hospital services are recorded on statutorily defined datasets, known as the Secondary Uses Service (SUS). The SUS data are cleaned and combined on a national basis to create HES data. The HES contains administrative hospital data for all inpatient, outpatient and accident and emergency (A&E) admissions in England. And, they hold admission, clinical, utilisation and demographics details in format of episodes and spells (HSCIC, 2016a).

In this research, performances of the PARR, the CPM and the CPM update were used as the benchmark, since these tools use the HES data and are still being used by commissioners across England. These decision support tools help to rank and group patients based on anticipated intervention level, including case management, disease management, supported care, prevention and wellness promotion.

Most existing decision support tools based on hospital administrative data use logistic regression or Coxian Phase-type Distribution models (Paton et al., 2014; Kansagara et al., 2011; Lewis et al., 2011; DH, 2011; ACI, 2014; Bardsley, 2012; Bottle et al., 2014; Mesgarpour et al., 2016b; Adeyemi et al., 2013). Although these models are simple and popular, they have limited power, because of

algorithm shortfalls, restricted assumptions and weak variable selection strategies. In the area of healthcare risk modelling research, there have been many successful implementations of machine learning methods (Green et al., 2006; Nilsson et al., 2006; Song et al., 2004; Peelen et al., 2010; Lee et al., 2012). However, few studies used a Bayesian approach to address emergency hospital readmission problems (Álvaro-Meca et al., 2012; Demir and Chaussalet, 2011; Cui et al., 2015; Helm et al., 2015; Gupta et al., 2014; Huws et al., 2008).

This study develops an ensemble generative risk model of emergency readmission within a year to hospitals in England. The machine learning ensemble method is a powerful technique, which uses a finite set of weaker models and an algorithm to combine and optimise the performance of the ensemble model. The HES inpatient data was extracted from English hospitals and maintained by the Health and Social Care Information Centre (HSCIC) (HSCIC, 2016c). Based on a preprocessing framework (Mesgarpour et al., 2016a), features were cleaned, generated, filtered and ranked. Thereafter, a number of sub-models based on population characteristics were trained using a Bayes Point Machine (BPM) approach. Afterwards, an optimised ensemble model of these sub-models was generated. The proposed model, the Ensemble Risk Model of Emergency Admissions (ERMER), was trained, tested and validated using three different time-frames.

The paper is structured as follows. Firstly, we describe the data and then the process of selecting a minimal number of features. Thereafter, the applied BPM algorithm is defined and the ensemble model is presented. Finally, we discuss the results of training, testing and benchmarking the ERMER, against the CPM (DH, 2006), the PARR (Billings et al., 2006) and the CPM update by Billings et al. (2013) (Billings et al., 2013) models.

2. Methods

2.1. Data

Administrative databases are used to monitor healthcare systems in the UK, the USA and other countries. Furthermore, healthcare data, such as inpatient, A&E, outpatient and GP records are used in predictive modelling problems (Jensen et al., 2012; Mullins et al., 2006). In addition, clinical databases compliment administrative databases, but they are expensive and not usually open to

the public. According to a study (Raftery et al., 2005), the cost per record for clinical data can range from £10 to £60, compared to £1 per record for the HES database.

In this research, only the HES inpatient data was used. The available snapshot of the database includes records from April 1995 to April 2010. The inpatient table consists of 206,528,432 episodes. This excludes 39,403 episodes with invalid *admidate* (admission date) and 11,212,871 episodes with invalid *hesid* (patient ID). In addition, similarly to the PARR model, each sample covers about 20% of unique patients within the trigger year of the selected time-frame (Table 1).

Table 1
Selected samples from the HES Inpatient database.

Samples	Timeframe	Population size		Sample size		Filtered patients		
		Episodes	Patients	Episodes	Patients	Total	No prior spell	No post spell
Sample-1	1999/04 - 2004/03	18,885,777	7,206,133	6,347,067	1,441,227	1,157,873	492,458	148,950
Sample-2	2004/04 - 2009/03	31,731,488	8,104,748	11,394,152	1,615,347	1,410,923	395,522	110,961
Sample-3	2000/04 - 2005/03	32,217,541	7,370,830	6,449,169	1,474,166	1,324,712	671,919	194,097

Before the modelling stage, four stages of data preprocessing were carried out (Mesgarpour et al., 2016a). Firstly, the extracted data was sorted by patients and the order of episodes. Then, invalid records were excluded. Thereafter, several corrections and imputations were carried out on dates, Healthcare Resource Groups (HRG) and demographics. Finally, some of the continuous features were converted into discrete to better capture non-linear interactions with other features. And, some of the discrete features were categorised into bigger groups to reduce sparseness and overfitting risks.

Similarly to the PARR model, the data was divided into three years of prior history, one year of trigger admission and one year of prediction period (time horizon). Then, half of each sample was used for training (train sub-sample) and the rest was used for testing (test sub-sample). Furthermore, spells were grouped into superspells based on the admission dates. A patient superspell is as a unit of care for the patient, which is the combination of all same-day episodes by any provider.

In this study, different combinations of the train sub-samples and the test sub-samples were used, but train and test have fixed definitions throughout the analyses (Table 4). The train sub-samples are used for training, learning-curve and complexity analysis. The test sub-samples are used for testing, cross-validation and benchmarking. Furthermore, no separate validation sub-sample is defined, since different modelling methods are not being compared.

2.2. Features

Based on previous studies (Billings et al., 2006, 2013, 2012; NHS, 2011; Mullins et al., 2006; Bardsley et al., 2013) and additional exploratory analyses, four main groups of features were initially generated from the inpatient database: three years cross-sectional, one year cross-sectional, 90 days cross-sectional and trigger-point features. In total, 738 summary features were generated, which the main categories are presented in Table 2.

Table 2
Main categories of all the initially defined features.

Category	Sub-category
Administrative	Admission: patient classification; number of episodes and spells; admission, readmission and discharge times; source and methods of admission and discharge. Bed days: duration of spells; preoperative and post-operative durations. Geographical: provider code; region of treatment. ID: patient identification, and admission timeframe number. Speciality: speciality of consultant; palliative cares. Waiting time: admission waiting time.
Clinical	Diagnosis: Charlson comorbidity groups; Elixhauser comorbidity groups; frequent categories of diagnoses; Charlson comorbidity index version that is developed by Dr Foster unit (Aylin et al., 2010; Bottle et al., 2011) and adapted by the HSCIC (HSCIC, 2016b); PARR's HRGs *reference* conditions, using version 3.5. Operation: operation groups; number of operations; frequent categories of operations.
Patient	Demographic: age; deprivations; ethnicity; gender.

Usually, Kernel classifiers, such as the BPM and the Support Vector Machine (SVM), are resistant to over-fitting, because of a weight regularisation implementation (Cawley and Talbot, 2007, 2010). However, since the number of generated features was very high, a feature reduction strategy was needed. Based on the framework developed in the previous stage of our research (Mesgarpour et al., 2016a), four steps of feature filtering were carried out, in order to reduce the number of features and to better capture the underlying structure.

Initially, highly stationary features were removed (constant count $\geq 95\%$). Then, features that were highly linearly correlated were excluded (linear correlation coefficient $\geq 80\%$). Thereafter, based on the average importance, initially, the three-year cross-sectional features were included, and then other features were added. Next, the features were sorted based on importance across train subsamples using two different methods: a random-forest importance score and an SVM importance ranking. Finally, a step-wise BPM procedure was developed using a forward-selection approach (micro average precision $\geq 0.01\%$).

The applied random-forest algorithm is a non-linear method and is an implementation of Breiman's algorithm (Breiman, 2001), which applies significance test criteria (Hothorn et al., 2010). It performs recursive univariate splitting and selects covariates based on the significance test. The sig-

nificance test approach, unlike the maximising information, does not suffer a systematic tendency towards covariates with many possible splits or many missing values. However, highly similar features and linearly correlated features were excluded in the prior step, because the applied algorithm is sensitive to correlated features.

Moreover, the SVM Recursive Feature Extraction (SVM-RFE) algorithm proposed by Guyon (Guyon et al., 2002) is applied to rank features recursively using SVM. The SVM-RFE algorithm ranks the features by training an SVM with a linear kernel and removing the features with the smallest ranking criterion.

2.3. Modelling Approach

Logistic regression, neural network, decision trees, Bayesian models and kernel methods, such as SVM and Gaussian processes, are often used in healthcare data mining. In this research, the Bayes Point Machines (BPM) method was chosen, since it is not prone to overfitting, highly efficient in approximating the Bayesian average classifier.

BPMs (Herbrich et al., 2001; Minka, 2001a) are a type of nonlinear classification algorithm, that identify an average classifier known as a Bayes Point in a version space. A version space can be defined as a set of hypotheses, each of which is an approximation of the main hypothesis class. Similar to SVMs, BPMs are more geometrically motivated and they try to find a hyperplane with an optimal margin between classes. In contrast, logistic regression maximises the probability of data by optimising the distance of each point to the decision boundary.

The soft margin SVM can be thought of as an approximation to BPMs (Herbrich et al., 2001). SVMs (Vapnik and Vapnik, 1998) use a mapping to indirectly transform data into higher dimensional space using a kernel function. Then, they use quadratic programming to optimise the classification's hyperplanes using support vectors and margins. However, the complexity of SVMs are characterised by the number of support vectors, and are only efficient for a symmetric version space.

On the other hand, BPMs sample the Bayesian posterior (Eq. 1) for a nonlinear classification in a kernel space. Then, they approximate the centre of the version space, which is a set of consistent hypothesis, and the effective size is determined from the training sample. BPMs minimise the generalisation error over a set of hypotheses according to a prior probability, instead of maximising

the classification boundary margin explicitly, as SVMs do. The predictive distribution can be thought of as a linear discriminant function, which is assumed to have the following parametric density:

$$p(y|x,w) = p(y|s = w^T x) \tag{1}$$

where w is the weight or latent parameter vector, x is the fully observed feature vector, and s is the score function. BPMs use the kernel trick to find an optimised w, and the centre mass of the version space is approximated using an average of the weight vectors while minimising the average generalisation error. The derived scores are subject to additive Gaussian noise to allow for measurement or labelling errors (Eq. (2)).

$$\begin{aligned} p(y|s,\varepsilon) &= (ys + \varepsilon > 0)1 \\ \text{, with } p(\varepsilon) = N(\varepsilon|0,1),\ \wedge 1(\alpha > 0) &= \begin{cases} 1 & if\ \alpha > 0 \\ 0 & if\ \alpha \leq 0 \end{cases} \end{aligned} \tag{2}$$

In this research, Microsoft's Infer.Net library (Microsoft Research, 2016) was used to construct the BPM model. The applied algorithm uses the original version of the BPM, with two main modifications. Firstly, it uses a mixture of Gamma-Gamma, a heavy-tailed prior probability distribution for the precision of weights and features. Secondly, it applies the Expectation Propagation (EP) message passing to infer posterior probabilities, which has been demonstrated (Minka, 2001b,a) in Gaussian mixture problems to be better than approximation techniques. Therefore, it is invariant to parameter rescaling or shifting, unlike logistic regression or SVM. Moreover, active Bayesian training can allow continuous updates of the model and account for changes in the prior probabilities. Furthermore, the BPM can efficiently handle a relatively larger number of features.

2.4. Ensemble Model

Firstly, one main model (*cond_main*) and four conditional sub-models were specified with significantly diverse populations which represent unique clinical and behavioural categories (Fig. 1). The conditional sub-models includes: prior 12-month acute spells (*Cond_Prior-Acute-12-month*), prior

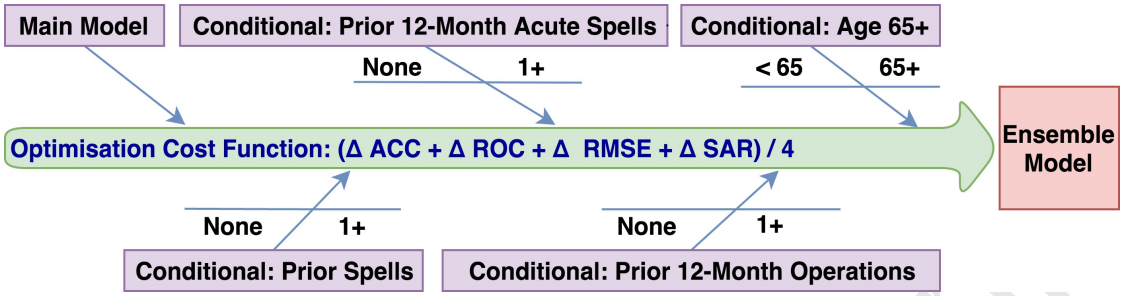


Fig. 1. The Ensemble model.

12-month operation (*Cond_Prior-Oper-12-month*), prior spells (*Cond_Prior-Spells*) and age 65+ (*Cond_Age-65p*).

Afterwards, they were trained and tested across the sub-sample combinations (Table 4). Considering that the filtered features are more relevant for the main model, the sub-models have very different performances but with stable weights.

Then, to improve the performance of the decision support system, we decided to use an ensemble model (Algorithm 1). Three main challenges in ensemble modelling were: method of constructing sub-classifiers, weighting the classifier and optimisation. Based on background research and multiple trials, a weighted average ranking method was constructed, in addition to a heuristic method to optimise the weights of sub-classifiers (Sewell, 2008; Rokach, 2010; Sammut and Webb, 2011; Zhou, 2012; Murphy, 2012).

In another word, the ERMER partitions the data instance space, based on some populations similarities (sub-models). Then, it uses *data envelop analysis* methodology (Charnes and Cooper, 1984) to assign weights to different classifiers (Rokach, 2005). In this research, we refer to this weight function as the cost function, because we applied a search technique to optimise the weights that are assigned to each sub-model.

The cost function for the optimisation was defined as a normalised combination of four performance metrics: ACC (Accuracy), AUC, RMSE (Root Mean Square Error) and SAR (Squared error, Accuracy and ROC area) (Brown, 2011; Alvarez, 2011; Fukunaga, 2013). The applied ensemble algorithm (Algorithm 1) uses a bidirectional hill-climbing algorithm with a greedy initial solution set ($models_{ensemble}$) to generate an optimised ensemble model from the sub-models.

Firstly, it generates an initial solution based on the main model and one other sub-model with the highest Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC). Then, a bidirectional hill-climbing (Russell and Norvig, 2002) heuristic was applied to optimise the average of the four performance metrics, through iterations, trials (*trials*) and across samples (*samples*).

The hill-climbing method is a greedy sequential search with forward and backward passes, where the *learning rate* for each performance metric can be tuned manually prior to the execution. The *learning rate* in the algorithm (Algorithm 1) defined using $alpha_{ensemble_{min}}$ for the performance indicators (Fukunaga, 2013; Caruana et al., 2004; Opitz and Maclin, 1999).

The sub-models in the ensemble heuristic are selected using a bagging ensemble (selection with replacement). Then, the sub-models are combined using a *mean* combiner, which is the approximate posterior probability based on the weighted average of the risk scores, without any additional training. When the first run of the algorithm, with the defined iterations, trails and train sub-samples, is finished; then, the second run, with less sensitive limits and thresholds, is executed using the best solutions of the first round.

$$\begin{aligned} model_{ensemble} = & Mean\{Cond\_Main + Cond\_Age\text{-}65p_0 + \\ & 9\ Cond\_Age\text{-}65p_1 + 4\ Cond\_Prior\text{-}Oper\text{-}12\text{-}month_0 + \\ & 2\ Cond\_Prior\text{-}Oper\text{-}12\text{-}month_1\}. \end{aligned} \tag{3}$$

Finally, the best performing ensemble model, with the minimum number of unique sub-models is selected. The optimised Ensemble Risk Models of Emergency Admissions (ERMER) based on our data sets is defined in Eq. (3). In this equation, a sub-model subscript represents the conditional state, and the coefficients represent the weights in the ensemble *mean* combiner.

3. Results

3.1. Goodness of fit

Four stages of performance checks were performed across test sub-samples to access the goodness of fit. Firstly, a learning-curve plot of training micro-average errors versus the number of training

points for sub-models was generated. The learning-curve is a function of the number of training points and the prediction accuracy rate, and it allows investigating the effect of sample sizes on the performance of models (Nordhausen, 2009; Murphy, 2012). Fig. 2a demonstrates that the train sub-sample size greater than 40,000 patients contributes very little to sub-models performances.

Table 3
The top significant features in the sub-models.

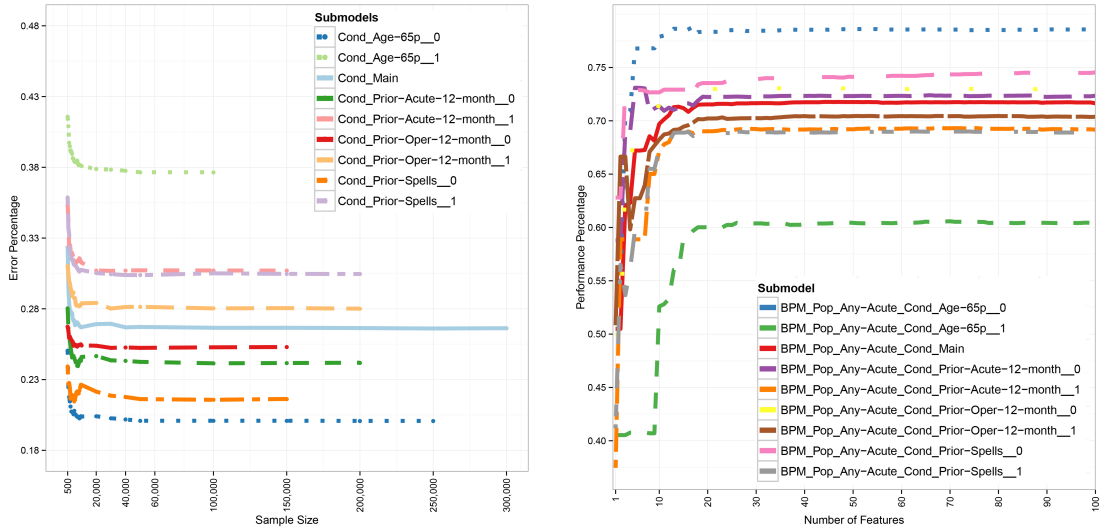
Feature	Calculation [a]
Sum of number of operations within 90 days and at the trigger.	$Count_{spell}(Unique_{spell}(opertn\_nn_{episode}))$
Count of recoded main speciality of state 'Maternity' in the past and at the trigger.	$mainspef_{spell} \in \{501, 560, 610\}$
Count of recoded main speciality of state 'Gynaecology' in the past and at the trigger.	$mainspef_{spell} \in \{502\}$
Count of recoded main speciality of state 'General' in the past and at the trigger.	$mainspef_{spell} \in \{300, 600, 620\}$
Having recoded gender of state 'Female'.	$sex_{patient} == 2$
Age of patient at the trigger.	$startage_{spell}$
Average of post-operative durations at the trigger.	$posopdur_{spell}$
Count of the acute admission method between 12 to 36 months, and within 90 days.	$Count_{spell}(admimeth_{spell} \in \{21, 22, 23, 24, 25, 2A, 2B, 2C, 2D, 28, 31, 32, 81, 82, 83, 84, 89, 98\})$
Average of spells durations in the past and at the trigger.	$Mean_{spell}(Max_{episode}(epidur))$
Average of gaps between admissions in the past.	$admidate_{spell_i} - dismeth_{spell_{i-1}}$
Having recoded ethnicity of state 'NA'.	$ethnos_{patient} \in \{S, 8, L, G\}$
Average value of the Charlson Index in the past.	$Mean(CharlsonIndex_{Dr\_Foster\_CCI}(diag\_nn_{spell}))$

[a] Refer the the HES dictionary for the definitions of the variables (HSCIC, 2010).

Moreover, the effects of complexity levels were investigated for the main model (*Cond_Main*) using F-score versus the number of features. The plot of the effects of complexity levels shows how the step-wise addition of top features changes the prediction performance of a model. Fig. 2b shows that adding up to 18 features (Table 3) from the sorted selected features improves the model's performance significantly; however, the gains then become very small (on average 0.005 change in AUC percentage). The presented learning-curve plot and complexity plot are for *Sample-1* train sub-sample, although the results are consistent across all other time-frames.

Thereafter, the convergences of the sub-models were tested using an iterative fitting, using train sub-samples, in order to assess over-fitting and variations in convergence. Fig. 3a shows that after the first few iterations, all sub-models converge quickly and after 40 iterations, the weights differences become very small.

Furthermore, a *k-fold* cross-validation (Murphy, 2012) algorithm was implemented for all the three test sub-samples (Table 4). Each test sub-sample was split into five equal-sized random samples. Then, $K-1$ folds was used for training and the *K-th* fold was used for validation. The final performance was generated after the cross-validation cycled through all the K combinations of splits. Fig. 3b exhibits very small standard deviations in the accuracy, the mean of negative log-probability and the AUC for the sub-models' cross-validations.



(a) Learning-curve of sub-models, micro-average error versus number of training points.

(b) Complexity analysis of sub-models, the F-score versus number of features.

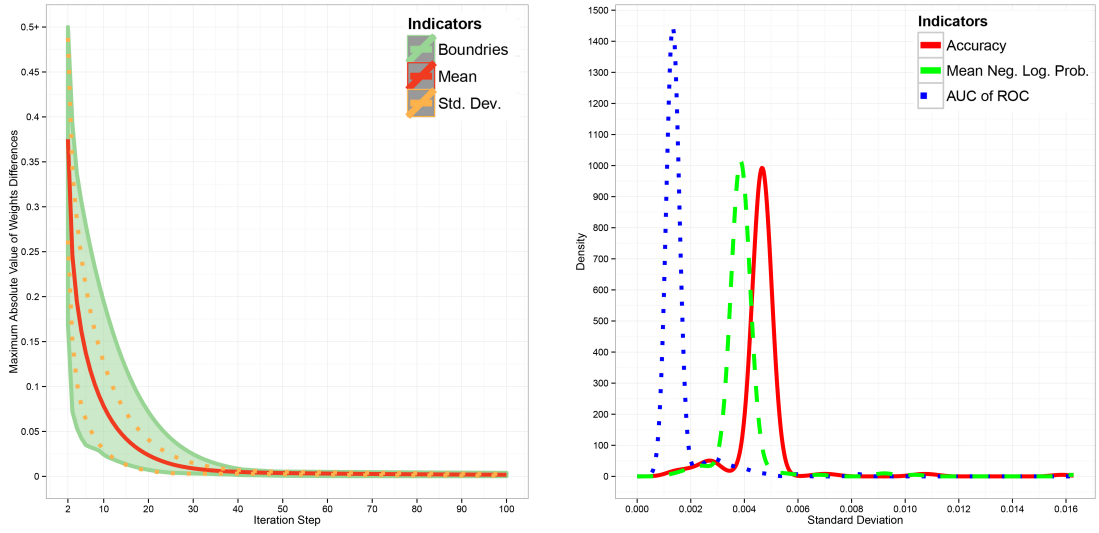
Fig. 2. Learning-curve and complexity analysis plots of sub-models (train sub-sample from Sample-1).

Finally, the profiling was done using the three test sub-samples, based on population characteristics and performance indicators (Table A.1 and Table A.2). Table A.3 demonstrates the weights of the features for each sub-model, as well as the features definitions, encoded categories and temporal states. In the following section the benchmark is discussed.

3.2. Benchmark

Admission risk models are limited by the characteristics of the selected subpopulation and data quality issues, such as missing diagnoses for outpatients and A&E patients (Billings et al., 2013), delayed death registration (ONS, 2014) and the number of registered or consented patients. Moreover, models developed by researchers usually have different settings and assumptions; hence, comparisons become more subjective.

The developed ERMER model is benchmarked against the CPM (DH, 2006; Paton et al., 2014), the PARR (Billings et al., 2006) and Billings et al. (2013) (Billings et al., 2013) models using the reported performance statistics.



(a) Average and range of convergence

(b) Distribution of standard deviation of cross-validation's performance indicators

Fig. 3. Summary statistics of convergence and cross-validation tests for all sub-models (test sub-samples).

For the testing, validation and benchmarking phase, three data settings were considered: *Sample-1*'s train and test sub-samples, *Sample-2*'s train and test sub-samples, and finally a rolling window setting with *Sample-1*'s train sub-sample and *Sample-3*'s test sub-sample (Table 4). The rolling window is configured as the one-year gap in admission trigger year, to better assess the stability of the model over time. In addition, for better comparison against the benchmark models, three different subpopulations were selected from the outputted test results (*Sub_PARR-2-Settings*, *Sub_IPAEOPGP* and *Sub_Any-Acute*).

Table 4

Combinations of test and train sub-samples.

Samples	Train Sub-Sample	Test Sub-Sample
Sample-1	*sample-1* training set	*sample-1* testing set
Sample-2	*sample-2* training set	*sample-2* testing set
Sample-1-train-half-3-test-half	*sample-1* training set	*sample-3* testing set

For comparison, numerical summaries beyond the ROC and abstract statistical summaries must be used to avoid misinterpretation (Steyerberg et al., 2010; Pencina et al., 2008; Cook, 2007). In addition to the ROC (Fig. 4), the profiling is presented using three forms of presentations: summary

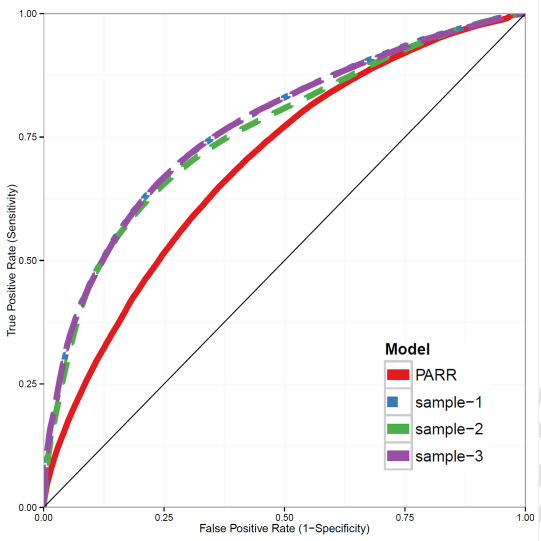


Fig. 4. ROC of the PARR model (reported figure) against the ERMER model (test sub-samples).

statistics for three risk cut-off points (Table 6) against the previous models (Table 5), summary statistics for 20 risk bands (Table A.1) and the profile of top risk segments (Table A.2).

The ERMER model made considerable improvement to the previous models. For instance, according to Table 6, the ERMER model with subpopulation *Sub_Any-Acute* has precision 0.719 and AUC of 0.771 with *Sample-1* as the test set, compared to 0.529, 0.73 for the Billings et al. (2013) model with inpatient (IP) data.

4. Discussion

In this study, a set of significant features was initially developed using a framework. Then, several predictive models were trained based on different subpopulations. The defined sub-models were fitted using a BPM algorithm, with Gamma priors, and EP message passing for the inference of the posterior. Furthermore, an optimised ensemble of five sub-models was produced based on the age group sub-models, the 1-year prior operation sub-models and the general model.

Table 5
The benchmark of the previous emergency readmission models (reported statistics)

Statistic	PARR			CPM	Billings-13 (IP)	Billings-13 (IPAEOPGP)[a]
Threshold[b]	0.50	0.60	0.70	0.50	0.50	0.50
The reported performance of the benchmarked models						
True & False Positive (TP+FP)	17,455	4,810	2,011	NR	8,743	10,545
True Positive (TP)	NR[c]	NR	NR	NR	4,627	5,669
Sensitivity (True Positive Rate)	0.543	0.178	0.081	NR	0.049	0.060
Specificity (True Negative Rate)	0.722	0.950	0.986	NR	NR	NR
Precision (Positive Predictive Value)	0.653	0.774	0.843	0.538	0.529	0.538
Emer. admi. post 12 m. per TP[d]	1.47	2.23	3.0	NR	NR	NR
Emer. admi. prior 12 m. per TP	2.22	3.43	4.59	NR	NR	NR
Emer. admi. prior 13-24 m. per TP	0.93	1.84	2.80	NR	NR	NR
Emer. admi. prior 25-36 m. per TP	0.73	1.48	2.25	NR	NR	NR
AUC of ROC	0.69			0.780	0.73	0.78
Total number of patients	42,778			281,617	1,836,099	1,836,099

[a] The Billings et. al. (2013) model with inpatient (IP), A&E (AE), outpatient (OP) and GP data.
[b] The threshold on the predicted risk.
[c] Not reported (NR).
[d] Average number of emergency readmission of the truly positively predicted patients.

Table 6
The benchmark of the ERMER model for different sub-populations using test sub-samples.

Statistic	Sub_PARR-2-Settings [a]			Sub_IPAEOPGP [b]			Sub_Any-Acute [c]		
Threshold	0.50	0.60	0.70	0.50	0.60	0.70	0.50	0.60	0.70
Train: train sub-sample of *Sample-1*; Test: test sub-sample of *Sample-1*									
True & False Positive (TP+FP)	19,646	7,946	2,991	51,422	30,361	14,719	52,842	31,260	15,231
True Positive (TP)	11,962	5,512	2,291	36,966	24,051	12,432	37,979	24,759	12,878
Sensitivity (True Positive Rate)	0.390	0.180	0.075	0.478	0.311	0.161	0.461	0.300	0.156
Specificity (True Negative Rate)	0.805	0.938	0.982	0.887	0.950	0.982	0.900	0.956	0.984
Precision (Positive Predictive Value)	0.609	0.694	0.766	0.719	0.792	0.845	0.719	0.792	0.846
Emer. admi. post 12 m. per TP	1.242	1.600	2.105	1.581	1.857	2.146	1.586	1.863	2.158
Emer. admi. prior 12 m. per TP	0.462	0.607	0.740	0.351	0.365	0.368	0.351	0.364	0.367
Emer. admi. prior 13-24 m. per TP	0.401	0.532	0.646	0.319	0.336	0.327	0.318	0.335	0.326
Emer. admi. prior 25-36 m. per TP	0.006	0.007	0.009	0.004	0.004	0.005	0.004	0.004	0.005
AUC of ROC	0.661			0.767			0.771		
Total number of patients	70,147			204,672			231,755		
Train: train sub-sample of *Sample-2*; Test: test sub-sample of *Sample-2*									
True & False Positive (TP+FP)	25,972	11,121	4,212	61,229	34,292	15,745	62,910	35,230	16,177
True Positive (TP)	15,916	7,577	3,169	43,858	26,920	13,180	45,032	27,611	13,539
Sensitivity (True Positive Rate)	0.470	0.224	0.094	0.503	0.309	0.151	0.492	0.302	0.148
Specificity (True Negative Rate)	0.745	0.910	0.974	0.873	0.946	0.981	0.883	0.950	0.983
Precision (Positive Predictive Value)	0.613	0.681	0.752	0.716	0.785	0.837	0.716	0.784	0.837
Emer. admi. post 12 m. per TP	1.296	1.604	2.051	1.623	1.925	2.272	1.624	1.922	2.270
Emer. admi. prior 12 m. per TP	0.452	0.591	0.723	0.365	0.403	0.441	0.365	0.402	0.440
Emer. admi. prior 13-24 m. per TP	0.388	0.507	0.635	0.327	0.361	0.395	0.327	0.360	0.393
Emer. admi. prior 25-36 m. per TP	0.007	0.009	0.010	0.005	0.006	0.007	0.005	0.006	0.007
AUC of ROC	0.663			0.756			0.759		
Total number of patients	73,315			224,001			243,712		
Train: train sub-sample of *Sample-1*; Test: test sub-sample of *Sample-3*									
True & False Positive (TP+FP)	22,351	8,351	2,896	60,515	35,642	18,487	62,213	36,753	19,117
True Positive (TP)	14,003	5,942	2,337	44,730	28,783	16,114	45,950	29,654	16,678
Sensitivity (True Positive Rate)	0.340	0.144	0.057	0.438	0.282	0.158	0.421	0.272	0.153
Specificity (True Negative Rate)	0.834	0.952	0.989	0.905	0.959	0.986	0.917	0.964	0.988
Precision (Positive Predictive Value)	0.627	0.712	0.807	0.739	0.808	0.872	0.739	0.807	0.872
Emer. admi. post 12 m. per TP	1.311	1.730	2.361	1.646	1.913	2.163	1.655	1.926	2.186
Emer. admi. prior 12 m. per TP	0.522	0.684	0.805	0.364	0.348	0.304	0.364	0.347	0.304
Emer. admi. prior 13-24 m. per TP	0.435	0.565	0.667	0.322	0.306	0.260	0.321	0.305	0.259
Emer. admi. prior 25-36 m. per TP	0.005	0.006	0.010	0.004	0.004	0.004	0.004	0.004	0.004
AUC of ROC	0.658			0.767			0.771		
Total number of patients	91,369			268,575			304,888		

[a] Population setting for the PARR-2 model: age: 65+; Trigger admission: Emergency.
[b] Population setting for the Billings et al. (2013) model: Age: 18-95; Trigger admission: Emergency.
[c] All the population for the selected sample: Trigger admission: Emergency admission.

Thereafter, the developed decision support tool, Ensemble Risk Model of Emergency Admissions (ERMER), was benchmarked against the PARR, the CPM and Billings et al. (2013) models, with very similar settings. The proposed model outperforms other models for any-emergency readmissions and the subpopulation of 18 to 95-year-old patients. The ROC of any-emergency readmission

is between 0.759 and 0.771, compared with the PARR, which is 0.69 with an age restriction (65+) and an HRG restriction (*reference* conditions). In addition, the performance is very close to the CPM and Billings et al. (2013) models, which predict any-emergency admissions using inpatient, outpatient, A&E and GP data.

4.1. Data

Firstly, the feature preparation is the most time-consuming part of many analyses. There are three main layers of difficulties in the preparation of features: correlations, recategorisations and selections (Mihaylova et al., 2011; Walpole et al., 2014; Yang et al., 2005). In this study, the variables were generated and selected based on the previously developed preprocessing framework (Mesgarpour et al., 2016a). Based on this framework, a large pool of variables was generated and reduced based on a set of defined criteria. Then, these were ranked and top features were inputted into the model.

Capturing high-risk patients using diagnoses can be difficult owing to variate coding practices, under-reporting of diagnostic variables, incomplete coding of transferred patients and comorbidities' complexity (Bottle et al., 2011; Billings et al., 2013; Reimer et al., 2016). Therefore, only high-level diagnoses groups were included and the remaining detailed codes were aggregated.

In this study, a recent version of Charlson index was used, which is actively maintained by the HSCIC and Dr Foster unit (Aylin et al., 2010; Bottle et al., 2011). Comorbidity scoring is usually used to distinguish the conditions present on admission from complications. But, poor coding and disregarding the effects of population characteristics can introduce bias (constant risk fallacy) (Nicholl, 2007; Fischer et al., 2011). Other criticisms of scoring originate from choosing small cohorts, using additive risk models of different medical conditions, ignoring important factors, such as the length-of-stay and the presence of different valid principal diagnoses across different cohorts (Quan et al., 2005; Bottle and Aylin, 2011).

Moreover, left-censored and right-censored observations introduce bias in the features and predicted risk estimates (Singer and Willett, 2003). According to Table 1, about 8% to 15% of patients do not have any admissions after the trigger event. In addition, about 28% to 51% of patients do not have any other prior admissions before the trigger event.

Finally, it has been speculated that many of the variations in readmission can be due to the delivery of the care method, which cannot be quantified using an administrative database only (Bottle et al., 2014; Billings et al., 2013; DH, 2006).

4.2. Model

There is always scepticism about machine learning because of the hypes or failures of inappropriate modelling approaches. For instance, Bottle et al. (2014) (Bottle et al., 2014) stated that machine learning methods, particularly Neural Networks (NN) and SVMs, did not offer noticeably better predictions for readmission risk compared to linear regression, and were relatively harder to implement. But, we believe there were four main possible flaws: missing influencing features in the Principal Component Analysis (Yang et al., 2005); using highly interdependent features, small training sets or a weak network design for the NN (Matignon, 2005); ignoring the temporal dimension and prior probabilities; and linearity and homogeneity assumptions (Congdon, 2010).

In general, accuracy and efficiency of a Bayesian model depend on five main design choices: the representation of features, fitness algorithm, inference approximation, assignment and update of prior probabilities, and the framework of system states.

Firstly, the features were carefully generated, selected and ranked before generating the models. The initial prototype models, without the aforementioned feature selection strategies, have shown very high sensitivity to intercorrelations, sparsity and noisy features. As a result, these caused non-convergence, weight decay and performance degradation.

Moreover, in comparison with the SVM, the BPM method is demonstrated (Herbrich et al., 2001) to provide a better solution for an asymmetric version space, to efficiently handle large datasets and to provide a smoother decision boundary.

Furthermore, Microsoft's version of the BPM algorithm (Microsoft Research, 2016) uses EP message passing, which in Gaussian mixture problems is demonstrated (Minka, 2001b,a) to be better than approximation techniques, such as the Markov Chain Monte Carlo, Laplace and Variational Bayes techniques. The EP does not guarantee convergence, but in practice in many cases, it does, especially if the features are not highly interdependent to become trapped in a region of local optima.

Finally, the choice of prior probability distributions of the weights and features can have a significant impact on the robustness of the algorithm. The applied algorithm uses a heavy-tailed prior, which is more robust towards outliers of the weight distributions. Also, the incremental Bayesian training of the ERMER allows it to incorporate the effects of changes in prior distributions.

4.3. Performance

All the sub-models are stable in the convergence and cross-validation testing. However, the features are initially selected based on the main model's population. The weights are very similar, proportionally, for all sub-models owing to very similar feature distributions, except for two: the sub-model with no prior spell ($Cond\_Prior\text{-}Spells_0$) and the sub-model with no prior operation ($Cond\_Prior\text{-}Oper\text{-}12\text{-}month_0$).

Furthermore, the applied BPM algorithm can handle a large number of features and a moderately large number of observations in comparison to logistic regression. On average, it takes about two to eight minutes [1] to train a sub-model with 100 features.

Also, the models performances are consistently high across all the test sub-samples. The performance of the main sub-models improves the ROC (Fig. 4), sensitivity, specificity and precision percentage by 2.83, 0.50, 1.26, and 2.83, respectively (Table 6).

Furthermore, the populations of readmitted patients are very low; therefore, the samples are significantly unbalanced in terms of the dependent variable. The main models have 3 to 4.5 times less, and sub-models have between 1 to 10 times less readmitted patients compared to non-readmissions. Therefore, based on the sensitivity, precision, and the ROC, models can more confidently identify low-risk patients, and avoid unnecessary interventions.

In addition, it improves the previous model (Mesgarpour et al., 2016a), which does not use the ensemble of subpopulations. The ROC and precision percentage of the any-acute model increase by 2.83 and 7.16, respectively, and sensitivity decreases in consequence.

Moreover, the features were selected based on the main model, which considers all the emergency admission population. Therefore, the PARR subpopulation under-performs. However, compared

[1] Windows 10 machine with Intel i7 2 GHz quad-core CPU and 8 GB 1600 MHz RAM.

to the PARR model, the predicted high-risk patients have less number of prior-admission for all the subpopulations, which makes it considerably harder to predict.

In addition, based on the population profile of the top 1000 risk segments (Table A.2), the model (*Any-Acute*) predicts more patients with chronic obstructive pulmonary disease (COPD), depression, diabetes, coronary heart disease (CHD), congestive heart failure (CHF) and smaller average age as high-risk, than the CPM and the PARR models did. On the other hand, cancer that is highly predictable and manageable has a smaller share among the high-risk patients.

Because sensitivity and precision vary across risk scores, and the costs of interventions or readmissions are not zero, it is better to define a profit function. However, owing to a lack of necessary variables for mapping costs, this was not considered

Finally, additional work is necessary to improve the comorbidity risk scoring and to dynamically adjust for temporal patterns.

5. Conclusion

In conclusion, the ERMER provides a generic approach in modelling readmission emphasising on robustness and feature discovery. Moreover, based on a large number of iterations for performance assessment across different settings, the ERMER maintained its high discriminatory performance. Consequently, the ERMER can bring a significant improvement to the current decision support system in use, improve care quality and reduce the costs.

Future research should aim to better adjust for comorbidity risk and temporal patterns.

Summary Points

What is already known?

- Avoidable emergency hospital admission can be an indicator of suboptimal care quality.
- Identification of high-risk patients for intervention can substantially improve care quality and reduce costs.
- Designing features and developing predictive models that can adjust continuously to a fast-changing health care system and population characteristics are very challenging.

What this paper adds?

- The optimised ensemble model of sub-populations was proved to increasingly improve the risk model.
- The combination of using a nonlinear Bayesian model and applying a preprocessing framework for feature generation and selection can effectively create a highly adaptable predictive model.
- The ensemble of generative models is a new effective way to predicts patients with harder predictability, such as patients with chronic conditions and patients with fewer prior hospitalisation records.

Authors' Contributions

MM preprocessed data, designed the model, drafted the manuscript and submitted the article. TC and SC provided valuable insights in the design and interpretation of results, revised the article critically and issued the final approval.

Acknowledgements

This work was supported by the HSCMG at the University of Westminster.

Conflicts of Interest

Authors do not have any conflicts of interest or financial interests to declare.

References

Adeyemi, S., Demir, E., Chaussalet, T., 2013. Towards an evidence-based decision making health-care system management: Modelling patient pathways to improve clinical outcomes. Decision Support Systems 55, 1, 117–125.

Alvarez, G. A., 2011. Representing multiple objects as an ensemble enhances visual cognition. Trends in cognitive sciences 15 (3), 122–131.

Álvaro-Meca, A., Kneib, T., Prieto, R. G., de Miguel, Á. G., 2012. Impact of comorbidities and surgery on health related transitions in pancreatic cancer admissions: A multi state model. Cancer epidemiology 36 (2), e142–e146.

Aylin, P., Bottle, A., Jen, M. H., Middleton, S., Intelligence, F., Nov. 2010. HSMR mortality indicators. [Retrieved 02.09.2016].
URL `https://www1.imperial.ac.uk`

Bardsley, M., Sep. 2012. Predictive models and data linkage - sharing international experience: Linking disease registry information and predictive modelling to improve quality and efficiency. [Retrieved 02.09.2016].
URL `http://www.nuffieldtrust.org.uk`

Bardsley, M., Blunt, I., Davies, S., Dixon, J., 2013. Is secondary preventive care improving? observational study of 10-year trends in emergency admissions for conditions amenable to ambulatory care. BMJ open 3 (1).

Bardsley, M., Georghiou, T., Chassin, L., Lewis, G., Steventon, A., Dixon, J., 2012. Overlap of hospital use and social care in older people in england. Journal of health services research & policy 17 (3), 133–139.

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). BMJ open 2 (4), e001667.

Billings, J., Dixon, J., Mijanovich, T., Wennberg, D., et al., 2006. Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. Bmj 333 (7563), 327.

Billings, J., Georghiou, T., Blunt, I., Bardsley, M., 2013. Choosing a model to predict hospital admission: an observational study of new variants of predictive models for case finding. BMJ open 3 (8), e003352.

Bottle, A., Aylin, P., 2011. Comorbidity scores for administrative data benefited from adaptation to local coding and diagnostic practices. Journal of clinical epidemiology 64 (12), 1426–1433.

Bottle, A., Gaudoin, R., Jones, S., Aylin, P., 2014. Can valid and practical risk-prediction or casemix adjustment models, including adjustment for comorbidity, be generated from english hospital administrative data (hospital episode statistics)? a national observational study. Health Serv Deliv Res 2 (40).

Bottle, A., Jarman, B., Aylin, P., 2011. Strengths and weaknesses of hospital standardised mortality ratios. BMJ 342, c7116.

Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32.

Brown, G., 2011. Ensemble learning. In: Encyclopedia of Machine Learning. Springer, pp. 312–320.

Caley, M., Sidhu, K., 2011. Estimating the future healthcare costs of an aging population in the UK: expansion of morbidity and the need for preventative care. Journal of Public Health 33 (1), 117–122.

Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A., 2004. Ensemble selection from libraries of models. In: Proceedings of the twenty-first international conference on Machine learning. ACM, p. 18.

Cawley, G. C., Talbot, N. L., 2007. Preventing over-fitting during model selection via bayesian regularisation of the hyper-parameters. The Journal of Machine Learning Research 8, 841–861.

Cawley, G. C., Talbot, N. L., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research 11, 2079–2107.

Charnes, A., Cooper, W. W., 1984. Preface to topics in data envelopment analysis. Annals of Operations research 2 (1), 59–94.

Clarke, A., Blunt, I., Bardsley, M., 2012. Ps18 analysis of emergency 30-day readmissions in england using routine hospital data 2004-2010. is there scope for reduction? Journal of Epidemiology and Community Health 66 (Suppl 1), A45–A45.

Congdon, P. D., 2010. Applied Bayesian hierarchical methods. CRC Press.

Cook, N. R., 2007. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 115 (7), 928–935.

Cui, Y., Torabi, M., Forget, E. L., Metge, C., Ye, X., Moffatt, M., Oppenheimer, L., 2015. Geographical variation analysis of all-cause hospital readmission cases in winnipeg, canada. BMC health services research 15 (1), 129.

Demir, E., Chaussalet, T., 2011. Capturing the re-admission process: focus on time window. Journal of Applied Statistics 38 (5), 951–960.

Fischer, C., Anema, H. A., Klazinga, N. S., 2011. The validity of indicators for assessing quality of care: a review of the european literature on hospital readmission rate. The European Journal of Public Health.

Fukunaga, K., 2013. Introduction to statistical pattern recognition. Academic press.

Green, M., Bjork, J., Forberg, J., Ekelund, U., Edenbrandt, L., Ohlsson, M., 2006. Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room. Artificial intelligence in medicine 38 (3), 305–318.

Gupta, S. K., Rana, S., Phung, D., Venkatesh, S., 2014. Keeping up with innovation: A predictive framework for modeling healthcare data with evolving clinical interventions. In: Proceedings of the 2014 SIAM International Conference on Data Mining. pp. 235–243.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine learning 46 (1-3), 389–422.

Helm, J. E., Alaeddini, A., Stauffer, J. M., Bretthauer, K. M., Skolarus, T. A., 2015. Reducing hospital readmissions by integrating empirical prediction with resource optimization. Production and Operations Management.

Herbrich, R., Graepel, T., Campbell, C., 2001. Bayes point machines. The Journal of Machine Learning Research 1, 245–279.

Hothorn, T., Hornik, K., Strobl, C., Zeileis, A., 2010. Party: A laboratory for recursive partytioning. [Retrieved 02.09.2016].
URL http://cran.r-project.org

Huws, D. W., Cashmore, D., Newcombe, R. G., Roberts, C., Vincent, J., Elwyn, G., 2008. Impact of case management by advanced practice nurses in primary care on unplanned hospital admissions: a controlled intervention study. BMC Health Services Research 8 (1), 115.

Jensen, P. B., Jensen, L. J., Brunak, S., 2012. Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics 13 (6), 395–405.

Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., Kripalani, S., 2011. Risk prediction models for hospital readmission: a systematic review. JAMA: the journal of the American Medical Association 306 (15), 1688–1698.

Lee, E. K., Yuan, F., Hirsh, D. A., Mallory, M. D., Simon, H. K., 2012. A clinical decision tool for predicting patient care characteristics: patients returning within 72 hours in the emergency department. In: AMIA Annual Symposium Proceedings. Vol. 2012. American Medical Informatics Association, p. 495.

Lewis, G., Aug. 2011. PARR++ is dead: long live predictive modelling. [Retrieved 02.09.2016].
URL http://www.nuffieldtrust.org.uk

Lewis, G., Curry, N., Bardsley, M., 2011. Choosing a predictive risk model: a guide for commissioners in england. [Retrieved 02.09.2016].
URL https://www.primis.nottingham.ac.uk

Matignon, R., 2005. Neural network modeling using SAS enterprise miner. AuthorHouse.

ACI, Dec. 2014. Risk stratification - a discussion paper for NSW health's approach to risk stratification. [Retrieved 02.09.2016].
URL http://www.aci.health.nsw.gov.au

DH, Dec. 2006. Combined predictive model - final report and technical documentation. [Retrieved 02.09.2016].
URL http://www.kingsfund.org.uk

DH, Aug. 2011. Risk stratification and next steps with DH risk prediction tools: Patients at risk of re-hospitalisation and the combined predictive model. [Retrieved 02.09.2016].
URL https://www.gov.uk

DH, Apr. 2012. Long term conditions compendium of information (third edition). [Retrieved 02.09.2016].
URL https://www.gov.uk

DH, 2013. Business case: for the health and care modernisation transition programme. [Retrieved 02.09.2016].
URL https://www.gov.uk

HSCIC, Sep 2010. Hes data dictionary. [Retrieved 02.09.2016].
URL http://www.hscic.gov.uk

HSCIC, Dec. 2013. Hospital episode statistics, emergency readmissions to hospital within 28 days of discharge - financial year 2011/12. [Retrieved 02.09.2016].
URL http://www.hscic.gov.uk

HSCIC, Jan. 2016a. Hospital episode statistics (HES). [Retrieved 02.09.2016].
URL http://www.hscic.gov.uk

HSCIC, 2016b. Summary hospital-level mortality indicator. [Retrieved 02.09.2016].
URL http://www.hscic.gov.uk

HSCIC, 2016c. What HES data are available? [Retrieved 02.09.2016].
URL http://www.hscic.gov.uk

Microsoft Research, 2016. Infer.net software solution. [Retrieved 02.09.2016].
URL http://research.microsoft.com

NHS, Oct. 2011. Scottish patients at risk of readmission and admission (SPARRA) - version 3 (developing risk prediction to support preventative and anticipatory care in scotland). [Retrieved

02.09.2016].
URL http://www.isdscotland.org

NHS, 2013. NHS england publishes CCG funding allocations for next two years following adoption of new formula. [Retrieved 02.09.2016].
URL http://www.england.nhs.uk

NHS, Oct. 2014. Five year forward view. [Retrieved 02.09.2016].
URL https://www.england.nhs.uk

Nuffield Trust, Aug. 2012. Predicting risk of hospital readmission with PARR-30. [Retrieved 02.09.2016].
URL http://www.nuffieldtrust.org.uk

ONS, Jul. 2014. Quality and methodology information: Mortality statistics in england and wales. [Retrieved 02.09.2016].
URL http://www.ons.gov.uk

The King's Fund, 2016. Predicting and reducing re-admission to hospital. [Retrieved 02.09.2016].
URL http://www.kingsfund.org.uk

Mesgarpour, M., Chaussalet, T., Chahed, S., June 2016a. Risk modelling framework for emergency hospital readmission, using hospital episode statistics inpatient data. In: 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS). pp. 219–224.

Mesgarpour, M., Chaussalet, T., Worrall, P., Chahed, S., June 2016b. Predictive risk modelling for integrated care: A structured review. In: 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS). pp. 42–47.

Mihaylova, B., Briggs, A., O'Hagan, A., Thompson, S. G., 2011. Review of statistical methods for analysing healthcare resources and costs. Health economics 20 (8), 897–916.

Minka, T. P., 2001a. Expectation propagation for approximate bayesian inference. In: Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., pp. 362–369.

Minka, T. P., Sep. 2001b. A family of algorithms for approximate bayesian inference. [Retrieved 02.09.2016].
URL http://research.microsoft.com

Mullins, I. M., Siadaty, M. S., Lyman, J., Scully, K., Garrett, C. T., Miller, W. G., Muller, R., Robson, B., Apte, C., Weiss, S., et al., 2006. Data mining and clinical data repositories: Insights from a 667,000 patient data set. Computers in biology and medicine 36 (12), 1351–1377.

Murphy, K. P., 2012. Machine learning: a probabilistic perspective. MIT press.

Nicholl, J., 2007. Case-mix adjustment in non-randomised observational evaluations: the constant risk fallacy. Journal of epidemiology and community health 61 (11), 1010–1013.

Nilsson, J., Ohlsson, M., Thulin, L., Höglund, P., Nashef, S. A., Brandt, J., 2006. Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. The Journal of thoracic and cardiovascular surgery 132 (1), 12–19.

Nordhausen, K., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition by Trevor Hastie, Robert Tibshirani, Jerome Friedman. Blackwell Publishing Ltd.

Opitz, D., Maclin, R., 1999. Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research, 169–198.

Paton, F., Wilson, P., Wright, K., 2014. Predictive validity of tools used to assess the risk of unplanned admissions: A rapid review of the evidence. [Retrieved 02.09.2016].
URL https://www.york.ac.uk

Peelen, L., de Keizer, N. F., de Jonge, E., Bosman, R.-J., Abu-Hanna, A., Peek, N., 2010. Using hierarchical dynamic bayesian networks to investigate dynamics of organ failure in patients in the intensive care unit. Journal of biomedical informatics 43 (2), 273–286.

Pencina, M. J., D'Agostino, R. B., D'Agostino, R. B., Vasan, R. S., 2008. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Statistics in medicine 27 (2), 157.

Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J.-C., Saunders, L. D., Beck, C. A., Feasby, T. E., Ghali, W. A., 2005. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Medical care, 1130–1139.

Raftery, J., Roderick, P., Stevens, A., 2005. Potential use of routine databases in health technology assessment. Health Technology Assessment 9 (20), 1–106.

Reimer, A. P., Milinovich, A., Madigan, E. A., 2016. Data quality assessment framework to assess electronic medical record data for use in research. International journal of medical informatics 90, 40–47.

Rokach, L., 2005. Ensemble methods for classifiers. In: Data mining and knowledge discovery handbook. Springer, pp. 957–980.

Rokach, L., 2010. Pattern classification using ensemble methods. Vol. 75. World Scientific.

Russell, S. J., Norvig, P., 2002. Artificial intelligence: a modern approach (International Edition). {Pearson US Imports & PHIPEs}.

Sammut, C., Webb, G. I., 2011. Encyclopedia of machine learning. Springer Science & Business Media.

Sewell, M., 2008. Ensemble learning. RN 11 (02).

Singer, J. D., Willett, J. B., 2003. Applied longitudinal data analysis: Modeling change and event occurrence. Oxford university press.

Song, X., Mitnitski, A., Cox, J., Rockwood, K., 2004. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. Stud Health Technol Inform 107 (Pt 1), 736–40.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., Kattan, M. W., 2010. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass.) 21 (1), 128.

Timmins, N., may 2013. Never again? the story of the health and social care act 2012.

Vapnik, V. N., Vapnik, V., 1998. Statistical learning theory. Vol. 1. Wiley New York.

Walpole, R., Myers, R., Myers, S., Ye, K., 2014. Probability and Statistics for Engineers and Scientists. Pearson.

Yang, Y., Webb, G. I., Wu, X., 2005. Discretization methods. In: Data mining and knowledge discovery handbook. Springer, pp. 113–130.

Zhou, Z.-H., 2012. Ensemble methods: foundations and algorithms. CRC press.

Algorithm 1 The ensemble modelling algorithm

Require: ▷ Set samples, modelling groups and sub-models

$samples \leftarrow$ {"*sample-1*","*sample-2*","*sample-3*","*sample-1-train-2-test*","*sample-1-train-3-test*"}
$groups \leftarrow$ {"*Pop_Any-Acute*","*Pop_Any-Acute-NO-Mental*"}
$models \leftarrow$ {"*Cond_Main*","*Cond_Spells*","*Cond_Acute_12_month*","*Cond_Oper_12_month*","*Cond_Age_65p*"}
$weight_{ensemble_{sum_{min}}} \leftarrow 20$ ▷ Second iteration of the algorithm: 300
$weight_{ensemble_{max}} \leftarrow 15$ ▷ Second iteration of the algorithm: 150
$search_{trials_{max}} \leftarrow 40$ ▷ Second iteration of the algorithm: 20
$search_{iterations_{max}} \leftarrow 150$ ▷ Second iteration of the algorithm: 150
$alpha_{ensemble_{min}} \leftarrow 0.0005$ ▷ Second iteration of the algorithm: 0.0001
$alpha_{model_{min}} \leftarrow 0.50$
TP_x : True positive of model x with cut-off point $alpha_{model_{min}}$
FP_x : False positive of model x with cut-off point $alpha_{model_{min}}$
FN_x : False negative of model x with cut-off point $alpha_{model_{min}}$
TPR_x : True positive rate of model x with cut-off point $alpha_{model_{min}}$
FPR_x : False positive rate of model x with cut-off point $alpha_{model_{min}}$

procedure EnsembleModels($models$) ▷ The ensemble modelling algorithm
  $models_{ensemble} \leftarrow$ InitialSolution($models$) ▷ Set the initial greedy solution
  $selected_{model} \leftarrow$ MainSearch($models_{ensemble}$) ▷ Run the main heuristic search

procedure InitialSolution($models$) ▷ The initial solution of the heuristic
  $model \in models$
  $model_{ensemble} \in models$
  $model_{max_{AUC}} \leftarrow MAX_{model_{AUC}}(models),\ WHERE\ model \neq$ "*Cond_Main*"
  return {"*Cond_Main*"} $\cup\ model_{max_{AUC}}$

procedure ACC(x) **return** $\frac{TP_x+TN_x}{TP_x+TN_x+FP_x+FN_x}$ ▷ ACC (Accuracy)

procedure AUC(x) **return** $\int_{-\infty}^{+\infty} TPR_x FPR'_x dx$ ▷ AUC of ROC

procedure RMSE(x) **return** $\sqrt{\frac{\sum_{i=1}^{n}(\hat{Y}_x - Y_x)}{n}}$ ▷ RMSE (Root Mean Square Error)

procedure SAR(x) **return** $\frac{ACC_x+ROC_x+(1-RMSE_x)}{3}$ ▷ SAR (Squared error, Accuracy, & ROC area)

procedure MainSearch($models_{ensemble}$) ▷ the main heuristic search
  $models_{selected} = \{\}$
  for all $s \in samples$ **do** ▷ Run for each sample
    for all $g \in groups$ **do** ▷ Run for each modelling group
      for $t \leftarrow 1, search_{trials_{max}}$ **do** ▷ Run trials
        for all $model_{ensemble} \in models_{ensemble}$ **do** ▷ Run for each initial solution
          $acc_0 \leftarrow auc_0 \leftarrow rmse_0 \leftarrow sar_0 \leftarrow 1$
          for $i \leftarrow 1, search_{iterations_{max}}$ **do** ▷ Run iterations
            $acc_i \leftarrow$ ACC($model_{ensemble}$)
            $auc_i \leftarrow$ AUC($model_{ensemble}$)
            $rmse_i \leftarrow$ RMSE($model_{ensemble}$)
            $sar_i \leftarrow$ SAR($model_{ensemble}$)
            $improvement \leftarrow (acc_i - acc_{i-1} >= alpha_{ensemble_{min}}) + (auc_i - auc_{i-1} >= alpha_{ensemble_{min}}) + (rmse_i - rmse_{i-1} >= alpha_{ensemble_{min}}) + (sar_i - sar_{i-1} >= alpha_{ensemble_{min}})$
            $degradation \leftarrow (acc_i - acc_{i-1} < -alpha_{ensemble_{min}}) + (auc_i - auc_{i-1} < -alpha_{ensemble_{min}}) + (rmse_i - rmse_{i-1} < -alpha_{ensemble_{min}}) + (sar_i - sar_{i-1} < -alpha_{ensemble_{min}})$
            if $i == 1$ **then** $backwardStep \leftarrow$ True ▷ Select a step
            else
              if $backwardStep ==$ True **then**
                if $degradation < 0.5$ **then** $model_{ensemble} \leftarrow model_{ensemble} \cup selected_{model}$ ▷ Forward
                else ▷ Switch
                  $backwardStep \leftarrow$ False and $switchStep \leftarrow$ True
              else
                if $backwardStep ==$ False **then**
                  if $improvement \geq 0.5$ **then** ▷ Switch
                    $switchStep \leftarrow$ True and $backwardStep \leftarrow$ True
              if $backwardStep ==$ True **then** ▷ Backward selection
                if $switchStep ==$ True **then** $counter \leftarrow 0$
                $model_{ensemble} \leftarrow model_{ensemble} \setminus model_{ensemble_{counter+1}}$
              if $backwardStep ==$ False **then** ▷ Forward selection
                $model \in models$
                $model_{ensemble} \leftarrow model_{ensemble} \cup model$
            $models_{selected} \leftarrow models_{selected} \cup model_{ensemble}$ ▷ Add selected model
  return $models_{selected}$

A. Additional Analyses Settings and Outputs

Table A.1

The risk bands statistics of the ERMER for different test sub-samples.

	PARR-2-Settings[a]						IPAEOPGP[b]						Any-Acute[c]					
Band	TP+FP	TP	Preci.	Sens.	Avg.[d]	C.I.[e]	TP+FP	TP	Preci.	Sens.	Avg.	C.I.	TP+FP	TP	Preci.	Sens.	Avg.	C.I.
Train: train sub-sample of *Sample-1*; Test: test sub-sample of *Sample-1*																		
1	14	0	0.000	0.000	0.00	0.00, 0.00	2,101	140	0.067	1.000	6.66	0.06, 0.08	3,797	240	0.063	1.000	6.30	0.06, 0.07
2	103	5	0.049	1.000	4.85	0.00, 0.10	8,065	945	0.117	0.871	11.68	0.11, 0.12	12,435	1,411	0.113	0.855	11.36	0.11, 0.12
3	522	48	0.092	0.906	9.19	0.07, 0.12	14,916	2,046	0.137	0.653	13.70	0.13, 0.14	20,067	2,675	0.133	0.618	13.33	0.13, 0.14
4	2,329	369	0.158	0.874	15.8	0.14, 0.17	15,054	2,643	0.176	0.458	17.55	0.17, 0.18	20,237	3,457	0.171	0.444	17.08	0.17, 0.18
5	3,404	742	0.218	0.637	21.79	0.20, 0.23	20,850	3,979	0.191	0.408	19.09	0.19, 0.20	24,368	4,613	0.189	0.372	18.92	0.18, 0.19
6	6,356	1,832	0.288	0.611	28.80	0.28, 0.30	20,969	4,585	0.219	0.320	21.87	0.21, 0.22	23,313	5,075	0.218	0.290	21.77	0.21, 0.22
7	7,681	2,618	0.341	0.466	34.09	0.33, 0.35	21,445	5,593	0.261	0.281	26.09	0.25, 0.27	23,063	5,968	0.259	0.255	25.87	0.25, 0.26
8	9,604	3,705	0.386	0.398	38.57	0.38, 0.40	18,623	6,271	0.337	0.239	33.64	0.33, 0.34	19,461	6,523	0.335	0.218	33.49	0.33, 0.34
9	11,501	5,080	0.442	0.353	44.18	0.43, 0.45	17,265	7,216	0.418	0.216	41.77	0.41, 0.43	17,827	7,425	0.417	0.199	41.66	0.41, 0.42
10	8,987	4,310	0.480	0.230	47.95	0.47, 0.49	13,962	6,896	0.494	0.171	49.38	0.49, 0.50	14,345	7,068	0.493	0.159	49.27	0.48, 0.50
11	6,913	3,713	0.537	0.166	53.66	0.53, 0.55	10,921	6,160	0.564	0.133	56.38	0.55, 0.57	11,191	6,313	0.564	0.124	56.38	0.55, 0.57
12	4,787	2,737	0.572	0.109	57.21	0.56, 0.59	10,140	6,755	0.666	0.127	66.57	0.66, 0.67	10,391	6,907	0.665	0.120	66.50	0.66, 0.67
13	3,076	1,948	0.633	0.072	63.32	0.62, 0.65	10,109	7,426	0.735	0.122	73.43	0.73, 0.74	10,357	7,585	0.732	0.116	73.21	0.72, 0.74
14	1,879	1,273	0.677	0.045	67.80	0.66, 0.70	5,533	4,193	0.758	0.065	75.81	0.75, 0.77	5,672	4,296	0.757	0.062	75.74	0.75, 0.77
15	1,116	800	0.717	0.027	71.68	0.69, 0.74	4,301	3,423	0.796	0.050	79.58	0.78, 0.81	4,424	3,517	0.795	0.048	79.49	0.78, 0.81
16	721	547	0.759	0.018	75.86	0.73, 0.79	2,975	2,447	0.823	0.035	82.31	0.81, 0.84	3,089	2,549	0.825	0.034	82.55	0.81, 0.84
17	460	364	0.791	0.012	79.13	0.75, 0.83	4,595	4,076	0.887	0.054	88.72	0.88, 0.90	4,757	4,223	0.888	0.053	88.77	0.88, 0.90
18	306	240	0.784	0.008	78.43	0.74, 0.83	1,697	1,475	0.869	0.019	86.91	0.85, 0.89	1,769	1,542	0.872	0.019	87.22	0.86, 0.89
19	199	167	0.839	0.005	83.92	0.79, 0.89	597	509	0.853	0.007	85.26	0.82, 0.88	619	527	0.851	0.006	85.13	0.82, 0.88
20	189	173	0.915	0.006	91.53	0.87, 0.95	554	502	0.906	0.006	90.61	0.88, 0.93	573	520	0.908	0.006	90.75	0.88, 0.93
N	70,147	30,671	0.609	0.390	43.72	0.43, 0.44	204,672	77,280	0.719	0.478	37.75	0.38, 0.38	231,755	82,434	0.719	0.461	35.56	0.35, 0.36

[a] The performance of the model for the sub-population *Sub_PARR-2-Settings*.

[b] The performance of the model for the sub-population *Sub_IPAEOPGP*.

[c] The performance of the model for the sub-population *Sub_Any-Acute*.

[d] The average of number of readmitted patients.

[e] The confidence interval for the average of number of readmitted patients using the bootstrapped central estimate with 95% CI.

Table A.2

The top risk segments profile of the predicted high-risk patients across test sub-samples.

Risk Seg.[a]	Model	Sub-population	Min Risk[b]	Asthma [c]	COPD [d]	Depres. [e]	Diab. [f]	Hyper. [g]	Cancer [h]	CHD [i]	CHF [j]	Avg. Age[k]	Avg. LoS[l]	5-9 Meds[m]	10+ Meds[n]
Train: train sub-sample of *Sample-1*; Test: test sub-sample of *Sample-1*															
		PARR-2-Settings	0.576	16.69	35.68	41.94	23.49	53.20	19.65	50.93	39.82	80.80	11.06	NA	NA
10,000		IPAEOPGP	0.759	11.54	12.19	12.24	8.25	20.71	6.51	14.88	10.34	39.68	4.49	NA	NA
		Any-Acute	0.766	11.25	11.52	11.57	7.93	19.74	6.24	14.01	9.72	38.61	4.39	NA	NA
		PARR-2-Settings	0.647	20.84	44.14	45.10	26.00	56.28	21.24	57.00	45.10	80.33	11.32	NA	NA
5,000		IPAEOPGP	0.817	15.80	15.78	15.28	10.14	25.20	7.52	18.58	12.96	42.36	4.91	NA	NA
		Any-Acute	0.818	15.84	15.72	15.32	10.38	25.12	7.60	18.48	12.92	41.99	4.93	NA	NA
		PARR-2-Settings	0.815	31.40	59.10	50.70	26.90	61.70	22.90	66.40	53.30	78.95	10.04	NA	NA
1,000	ERMER	IPAEOPGP	0.910	33.40	35.70	30.50	21.90	39.70	14.30	38.60	26.30	53.38	6.98	NA	NA
		Any-Acute	0.912	33.20	34.80	29.90	21.80	39.00	14.40	37.50	25.40	52.21	6.85	NA	NA
		PARR-2-Settings	0.881	37.40	67.60	52.00	26.40	63.20	25.20	69.60	55.20	77.98	9.35	NA	NA
500		IPAEOPGP	0.957	38.20	38.60	34.80	25.20	42.80	14.20	43.20	27.40	54.49	7.43	NA	NA
		Any-Acute	0.958	37.80	37.60	33.80	25.00	41.20	14.00	41.80	26.80	52.95	7.37	NA	NA
		PARR-2-Settings	0.933	36.40	70.00	53.60	27.60	63.20	25.60	69.60	53.60	77.34	9.57	NA	NA
250		IPAEOPGP	0.985	40.80	39.20	36.40	27.20	42.80	11.60	42.00	29.60	53.88	7.99	NA	NA
		Any-Acute	0.986	40.40	38.80	36.40	27.60	42.40	11.60	41.20	28.80	52.76	7.89	NA	NA

[a] The top predicted risk segment.

[b] The minimum predicted risk in the segment.

[c] The percentage of patients with a history of Asthma diagnosis (ICD-10: J45-J46).

[d] The percentage of patients with a history of Chronic Obstructive Pulmonary Disease (COPD) diagnosis (ICD-10: J20, J41-J44, J47).

[e] The percentage of patients with a history of Depression diagnosis (ICD-10: I10-I15).

[f] The percentage of patients with a history of Diabetes diagnosis (ICD-10: E10.0, E10.1, E10.6, E10.8, E10.9, E11.0, E11.1, E11.6, E11.8, E11.9, E12.0, E12.1, E12.6, E12.8, E12.9, E13.0, E13.1, E13.6, E13.8, E13.9, E14.0, E14.1, E14.6, E14.8, E14.9, E10.2-E10.5, E10.7, E11.2-E11.5, E11.7, E12.2-E12.5, E12.7, E13.2-E13.5, E13.7, E14.2-E14.5, E14.7).

[g] The percentage of patients with a history of Hypertension diagnosis (ICD-10: I10-I15, I27, I6, I87.0, I87, I97, K76.6, H35.0, R03, O13, O14, O16, O10, G93.2, H40.0, P292, P293).

[h] The percentage of patients with a history of Cancer diagnosis (ICD-10: C00-D49).

[i] The percentage of patients with a history of Coronary Heart Disease (CHD) diagnosis (ICD-10: I20-I25).

[j] The percentage of patients with a history of Congestive Heart Failure (CHF) diagnosis (ICD-10: I09.9, I11.0, I13.0, I13.2, I25.5, I42.0, I42.5-I42.9, I43.x, I50.x, P29.0).

[k] The average age of patients at the trigger event.

[l] The average length of stay of patient at the trigger event.

[m] The percentage of patients with 5-9 medication prescription.

[n] The percentage of patients with 10+ medication prescription.

Table A.3

The average importance of features and average weights of features in sub-models.

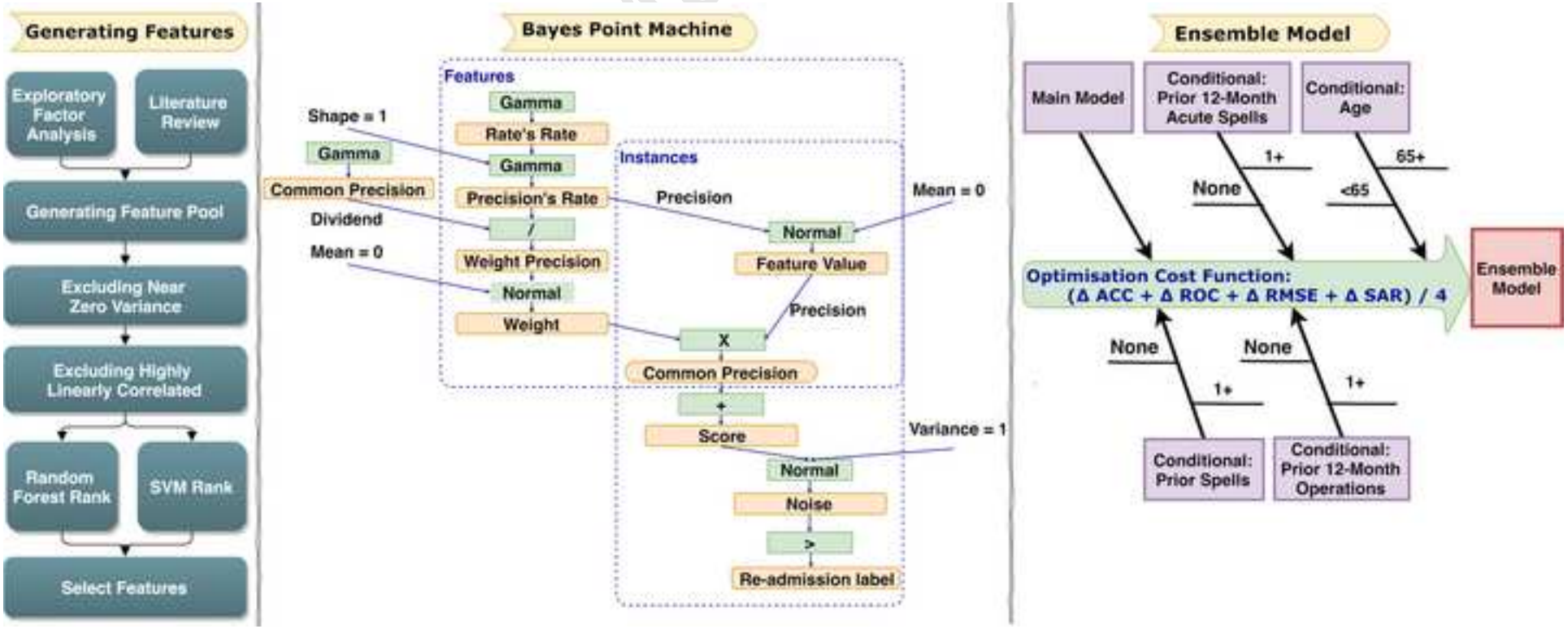
#	Feature	Sub-models:	Main	Age_0	Age_1	$Oper_0$	$Oper_1$
1	Sum of number of operations (trigger)		-0.744	<-1	-0.049	<-1	-0.666
2	Count of recoded main speciality of state 'Maternity' (trigger)		0.885	>1	-0.608	>1	0.289
3	Count of recoded main speciality of state 'Maternity' (3 years)		-0.021	-0.024	0.138	0.033	0.019
4	Count of recoded main speciality of state 'Gynaecology' (trigger)		0.742	0.910	-0.514	>1	0.311
5	Having recoded gender of state 'Female'		0.005	0.020	-0.032	0.059	-0.056
6	Count of recoded main speciality of state 'Gynaecology' (3 years)		-0.005	-0.004	-0.017	0.093	0.026
7	Age of patient (trigger)		0.023	0.003	0.004	0.015	0.024
8	Average of post-operative durations (trigger)		0.008	0.007	0.001	<-1	0.007
9	Count of the acute admission method (90 days)		0.054	0.053	0.043	0.044	0.049
10	Average of spells durations (3 years)		0.041	0.032	0.056	0.047	0.063
11	Sum of number of operations (90 days)		-0.011	-0.008	-0.002	>1	-0.004
12	Count of the acute admission method between (1-2 years)		-0.043	-0.031	0.075	-0.138	0.061
13	Count of recoded main speciality of state 'General' (trigger)		-0.077	-0.037	0.025	0.006	-0.183
14	Average of gaps between admissions (3 years)		0.236	0.189	0.180	0.375	0.154
15	Average of spells durations (trigger)		-0.002	-0.002	-0.002	-0.002	-0.001
16	Having recoded ethnicity of state 'others'		-0.363	-0.391	-0.368	-0.429	-0.320
17	Average of the Charlson Index (3 years)		0.018	0.040	0.009	0.011	0.018
18	Count of recoded main speciality of state 'General' (3 years)		0.015	0.021	0.003	0.013	0.012
19	Average of post-operative durations (3 years)		0.000	0.003	-0.001	>1	-0.002
20	Count of recoded main speciality of state 'General Surgery' (trigger)		-0.046	0.049	-0.064	-0.068	-0.145
21	Count of the acute admission method between 90 days to 12 months		-0.153	-0.152	-0.047	-0.314	-0.030
22	Average of pre-operative durations (trigger)		0.017	0.021	0.001	<-1	0.013
23	Count of recoded main speciality of state 'Plastic' (trigger)		0.025	0.161	-0.110	-0.164	-0.109
24	Having recoded ethnicity of state 'White'		0.010	-0.013	0.021	-0.003	0.015
25	Count of PARR's 'reference' conditions (90 days, trigger)		0.014	0.036	0.025	0.058	0.008
26	Count of recoded main speciality of state 'Geriatric' (3 years)		0.007	0.045	0.003	-0.011	0.006
27	Recoded Index of Multiple Deprivation Overall Rank (10 equal ranges)		-0.002	-0.006	0.002	0.013	-0.008
28	Maximum value of the Charlson Index (1 year)		-0.001	-0.009	0.002	-0.007	-0.003
29	Average of pre-operative durations (3 years)		-0.002	0.001	0.002	0.126	-0.004
30	Count of recoded main speciality of state 'General Surgery' (3 years)		0.018	0.024	0.002	-0.012	0.006
31	Count of recoded main speciality of state 'Plastic' (3 years)		0.002	0.013	-0.010	-0.001	-0.016
32	Count of external causes or complications (3 years)		0.007	0.003	0.003	0.020	-0.005
33	Count of recoded main speciality of state 'Geriatric' (trigger)		-0.017	-0.053	0.056	0.145	-0.136
34	Count of recoded main speciality of state 'A&E' (trigger)		-0.204	-0.182	0.013	-0.065	-0.335
35	Count of ischemic heart conditions (90 days, trigger)		-0.008	-0.012	0.002	0.003	-0.005
36	Count of unique main speciality seen (trigger)		0.119	0.150	0.029	-0.007	0.090
37	Average of post-operative durations (1 year)		-0.001	-0.002	-0.001	<-1	0.000
38	Count of other heart conditions (90 days, trigger)		0.016	0.011	0.009	0.013	0.007
39	Count of the elective admission method (90 days)		0.009	0.004	0.008	0.000	0.002
40	Count of thrombocytopenia, thrombocytosis & high WBC (90 days, trigger)		0.004	0.004	-0.002	-0.004	0.000
41	Count of recoded intended admission of states 'others' or 'Maternity' (90 days)		-0.016	-0.014	-0.006	-0.016	-0.008
42	Count of recoded main speciality of state 'A&E' (3 years)		0.032	0.037	0.034	0.074	0.037
43	Count of ACS respiratory conditions (90 days, trigger)		0.014	0.019	0.014	0.026	0.011
44	Count of ACS neurological disorders (90 days, trigger)		-0.014	0.001	-0.005	-0.019	-0.012
45	Count of mental conditions (90 days, trigger)		-0.001	0.052	-0.023	-0.003	0.003
46	Count of recoded main speciality of state 'Psychiatry' (3 years)		0.004	0.007	-0.013	0.019	-0.002
47	Count of recoded main speciality of state 'Psychiatry' (trigger)		-0.078	-0.029	0.058	0.035	-0.181
48	Count of the admission sources from 'others' or 'Maternity'		0.001	0.021	-0.012	-0.033	0.000
49	Count of chronic pulmonary conditions (trigger)		0.013	-0.039	0.027	0.034	0.018
50	Count of recoded main speciality of state 'Cardiothoracic' (3 years)		0.003	0.014	-0.013	0.047	-0.008
51	Count of ACS diabetes conditions (90 days, trigger)		0.013	0.022	0.010	0.012	0.013
52	Count of blood loss anemia conditions (90 days)		0.022	-0.019	0.008	0.032	0.019
53	Average of pre-operative durations (1 year)		0.003	0.001	-0.001	0.516	0.003
54	Count of recoded main speciality of state 'ENT' (trigger)		-0.080	-0.053	-0.127	-0.199	-0.144
55	Count of recoded region of state 'Eastern' (trigger)		0.455	0.534	0.583	0.593	0.220
56	Sum of number of operations between 90 days to 12 months		-0.020	-0.026	0.040	>1	-0.019

Continued on next page

#	Feature Sub-models:	Main	Age_0	Age_1	$Oper_0$	$Oper_1$
57	Count of organisation cluster of state 'Acute teaching trust' (trigger)	-0.029	-0.022	-0.030	-0.055	-0.020
58	Count of cardiac arrhythmias conditions (90 days, trigger)	0.002	-0.018	0.005	0.009	0.007
59	Count of congestive heart failure conditions (90 days)	-0.036	-0.046	-0.025	-0.030	-0.027
60	Count of recoded main speciality of state 'Ophthalmology' (3 years)	0.057	0.025	0.048	-0.084	0.033
61	Count of recoded main speciality of state 'Gastroenterology ' (3 years)	0.032	0.045	0.009	0.049	0.017
62	Count of ACS respiratory conditions (90 days, trigger)	-0.009	-0.015	-0.008	-0.005	-0.008
63	Count of recoded main speciality of state 'Cardiothoracic' (trigger)	0.084	0.174	0.081	0.081	-0.007
64	Count of organisation cluster of state 'Large acute trust' (trigger)	-0.058	-0.048	-0.035	-0.034	-0.052
65	Count of recoded main speciality of state 'ENT' (3 years)	0.033	0.038	0.043	-0.060	0.016
66	Count of recoded region of state 'Trent' (trigger)	0.506	0.592	0.602	0.654	0.284
67	Count of other neurological disorders (90 days)	0.007	-0.005	-0.002	0.018	0.008
68	Count of recoded region of state 'West Midlands' (trigger)	0.510	0.619	0.549	0.660	0.270
69	Count of recoded region of state 'London' (trigger)	0.511	0.604	0.616	0.627	0.291
70	Count of neoplasm conditions (90 days)	0.021	0.096	0.013	-0.039	0.026
71	Average of spells durations (1 year)	0.000	0.000	0.000	0.000	0.000
72	Count of recoded region of state 'North West' (trigger)	0.471	0.577	0.560	0.586	0.245
73	Count of the elective admission method between (1-2 years)	0.011	0.012	0.051	0.064	0.007
74	Count of heart operations (90 days)	-0.005	-0.008	-0.015	>1	-0.015
75	Count of recoded main speciality of state 'Urology' (3 years)	0.023	0.018	0.025	0.062	0.014
76	Count of organisation cluster of state 'Medium acute trust' (trigger)	-0.040	-0.053	-0.019	-0.027	-0.049
77	Count of recoded region of state 'Northern and Yorkshire' (trigger)	0.503	0.604	0.572	0.634	0.279
78	Count of provider type of state 'Trust' (trigger)	-0.107	-0.148	-0.124	0.014	-0.073
79	Count of malignancy conditions, except malignant neoplasm of skin (90 days)	-0.019	-0.096	-0.003	0.016	-0.023
80	Count of recoded main speciality of state 'Respiratory' (3 years)	0.026	0.049	-0.001	0.042	0.013
81	Count of depression conditions (90 days)	-0.007	-0.010	0.004	-0.016	0.004
82	Count of recoded main speciality of state 'Urology' (trigger)	0.069	0.080	0.012	-0.134	-0.013
83	Count of provider type of state 'Trust' (3 years)	0.085	0.137	0.126	-0.004	0.076
84	Count of recoded main speciality of state 'Gastroenterology' (trigger)	-0.012	0.064	0.029	0.011	-0.079
85	Count of recoded region of state 'others' (trigger)	0.519	0.620	0.599	0.679	0.284
86	Count of zero waiting time for elective admissions (3 years)	0.002	0.005	-0.010	-0.008	0.001
87	Count of organisation cluster of state 'Small acute trust' (trigger)	-0.054	-0.066	-0.017	-0.017	-0.050
88	Count of recoded region of state 'South East' (trigger)	0.540	0.643	0.614	0.617	0.326
89	Count of recoded main speciality of state 'Respiratory' (trigger)	-0.052	0.001	0.051	-0.028	-0.110
90	Count of liver conditions (90 days)	-0.003	-0.004	0.010	-0.009	0.001
91	Count of urinary operations (90 days)	-0.005	0.018	-0.022	>1	-0.008
92	Count of exposure to tobacco smoke conditions (trigger)	0.008	0.009	0.024	0.020	0.010
93	Count of external causes of morbidity conditions (90 days)	-0.014	-0.007	-0.008	0.037	-0.009
94	Count of cerebrovascular (stroke) conditions (90 days)	-0.022	-0.027	-0.012	0.008	-0.020
95	Count of disorders of lipidemias conditions (90 days)	-0.035	-0.043	-0.036	0.005	-0.033
96	Count of influenza a pneumonia conditions (90 days)	0.024	-0.003	0.012	-0.004	0.019
97	Count of recoded main speciality of state 'Ophthalmology' (trigger)	-0.076	0.008	-0.363	0.013	-0.250
98	Count of other veins, lymphatics and lymph nodes conditions (90 days)	-0.010	-0.005	-0.019	-0.002	-0.007
99	Count of days gap from the previous spell (trigger)	-0.727	-0.695	-0.662	-0.840	-0.521

Highlights:

- Using a Bayes Point Machine method, which has no hyper-parameter and is adaptive to changes in prior distributions of features, to predict the risk of emergency readmission to hospitals in the English National Health Service.
- Using an ensemble model to improve the performance of risk prediction and allow sensitivity and precision to be adjusted based on a cost function.
- Using a framework to collect a pool of features.
- Using a minimal amount of administrative data to capture the underlying structure better.


Generating Features
Exploratory Factor Analysis
Literature Review
Generating Feature Pool
Excluding Near Zero Variance
Excluding Highly Linearly Correlated
Random Forest Rank
SVM Rank
Select Features
Bayes Point Machine
Features
Gamma
Rate's Rate
Gamma
Precision's Rate
/
Weight Precision
Normal
Weight
Shape = 1
Gamma
Common Precision
Dividend
Mean = 0
Instances
Precision
Normal
Feature Value
Precision
X
Common Precision
+
Score
Normal
Noise
>
Re-admission label
Mean = 0
Variance = 1
Ensemble Model
Main Model
Conditional: Prior 12-Month Acute Spells
Conditional: Age
1+
None
65+
<65
Optimisation Cost Function:
(Δ ACC + Δ ROC + Δ RMSE + Δ SAR) / 4
Ensemble Model
None
None
1+
1+
Conditional: Prior Spells
Conditional: Prior 12-Month Operations