A machine learning approach to intensive care discharge.

Christopher J. McWilliams[1], Daniel J. Lawson[2], Raul Santos-Rodriguez[1], Iain D Gilchrist, Alan Champneys[1], Timothy H. Gould[4], Matthew J.C. Thomas[4], and Christopher P. Bourdeaux[4]

1 Department of Engineering Mathematics, University of Bristol
2 Integrative Epidemiology Unit, Population Health Sciences, University of Bristol
3 School of Experimental Psychology, University of Bristol
4 Intensive Care Unit, Queens Building, University Hospitals Bristol

*Corresponding author: chris.mcwilliams@bristol.ac.uk

Keywords: patient discharge, machine learning, clinical decision support, critical care, patient flow

Word count: 3554 (main text only)

Abstract

Objective
The primary objective is to work towards a clinical decision support tool that can improve discharge practice on the intensive care unit.

Design
We used two datasets of routinely collected patient data to test and improve upon a set of previously proposed discharge criteria.

Setting
Bristol Royal Infirmary general intensive care unit (GICU).

Patients
Two cohorts derived from historical datasets: 1933 intensive care patients from GICU in Bristol, and 10658 from MIMIC-III (a publicly available intensive care dataset).

Interventions
None.

Primary outcome measure
None

Results
In both cohorts few successfully discharged patients met the of all the discharge criteria. Both a random forest and a logistic classifier, trained on MIMIC and cross validated on GICU, demonstrated improved performance over the original criteria and generalised well between the cohorts. The classifiers showed good agreement on which features were most predictive of readiness-for-discharge, and these were generally consistent with clinical experience. By weighting the NLD criteria according to feature importance from the logistic model we showed improved performance over the original NLD criteria, while retaining good interpretability.

Conclusions
Our findings constitute a proof of concept for a decision support tool to run alongside a clinical information system, and streamline the process of discharge from the ICU.

Strengths and Limitations of this study:

- This study applies machine learning techniques to the problem of classifying patients that are ready for discharge from intensive care.
- Two cohorts of historical data are used, allowing cross-validation and a comparison of results between healthcare contexts.
- Our approach represents the first step towards a decision support tool that would help clinicians identify dischargeable patients as early as possible.

Introduction

Demand for intensive care unit (ICU) beds is rising at a time when resource is constrained[1]. In order to optimise the allocation of this resource, patients should be discharged from the ICU as soon as soon as they no longer require the specialist input provided there. The reduced ICU capacity caused by discharge delay can result in the delayed admission of emergency patients requiring ICU care[2,3]. Furthermore, patients remaining in the ICU after they are medically fit to leave experience detrimental effects on physical rehabilitation and psychosocial well-being[4].

The identification of individuals that are ready to leave ICU is a key component of patient flow through the hospital. At present this identification is a manual process, relying on physicians reviewing patients on a ward round at a standard point in time. There is a lack of formal guidance to inform discharge readiness and as such the process is sensitive to both the decision making heuristics of individual clinicians and structural factors within the hospital[5]. These issues may introduce undesirable variability in the timeliness of the discharge decision. A number of studies have looked to address this problem by attempting to standardise the discharge process[6]. One such study by Knight[7] proposed a set of physiological criteria to enable nurses to identify patients that are fit for discharge from a high dependency unit (HDU), thereby expediting the discharge process.

Increasingly ICUs are using clinical information systems (CIS) to collect, store and display physiological data. The availability of such routinely collected patient data presents the opportunity to apply methods from data science, with the potential to transform healthcare in a number of ways[8,9]. Two particular avenues for development are the automation of simple tasks[10] and the implementation of decision support systems[11], both of which would reduce the cognitive load of clinicians and free up scarce resource for tasks that require human expertise. We believe that the ICU discharge process is one area of healthcare practice that could be significantly improved by such data driven approaches.

In this study we investigated the possibility of using routinely collected data to identify patients that are ready for discharge. We studied two historical cohorts. One cohort consisted of patients treated on the general intensive care unit at the Bristol Royal Infirmary, while the second consisted of patients selected from the MIMIC-III database[12] (see Materials and Methods for details). Since there is no generally accepted definition of readiness-for-discharge, we adopted the criteria published by Knight[7] as a baseline for automation. Given the rational behind these criteria they represent a general and highly conservative set of constraints on physiology that characterise a patient as suitable for care on an acute ward (level 1 care). We initially posed the question: how many patients met the proposed criteria at the time they were declared ready for discharge? We then attempted to improve upon the performance of the original criteria by using machine learning techniques to identify patients that were ready for discharge. This work represents the first steps towards a decision support tool that will run in real time alongside our CIS and help clinicians identify dischargeable patients.

Methods

Discharge criteria

The nurse-led discharge (NLD) criteria proposed by Knight[7] consist of a set of constraints on various routinely collected vital signs and laboratory results. If a patient meets all the constraints for a period of at least four hours, Knight states that they may be safely discharged by a nurse. In order to test the NLD criteria on historical patient data we codified the constraints (see online supplementary file section A) into 15 binary tests, which are defined in table 1.

Test ID	Test name	Variable	Test condition
R0	Respiratory: airway	airway	airway patent
R1	Respiratory: FiO2	fio2	fio2 $\leq$ 0.6
R2	Respiratory: blood oxygen	spo2	spo2 $\geq$ 95 (%)
R3	Respiratory: bicarbonate	hco3	hco3 $\geq$ 19 (mmol/L)
R4	Respiratory: rate	resp (rate)	10 $\leq$ resp $\leq$ 30 (bpm)
C0	Cardiovascular: blood pressure	bp (systolic)	bp $\geq$ 100 (mmHg)
C1	Cardiovascular: heart rate	hr	60 $\leq$ hr $\leq$ 100 (bpm)
P	Pain	pain	0 $\leq$ pain $\leq$ 1
CNS	Central nervous system	gcs	gcs $\geq$ 14
T	Temperature	temp	36 $\leq$ temp $\leq$ 37.5 (C)
B0	Bloods: haemoglobin	haemoglobin	haemoglobin $\geq$ 9 (g/dL)
B1	Bloods: potassium	k	3.5 $\leq$ k $\leq$ 6.0 (mmol/L)
B2	Bloods: sodium	na	130 $\leq$ na $\leq$ 150 (mmol/L)
B3	Bloods: creatinine	creatinine	59 $\leq$ creatinine $\leq$ 104 (umol/L)
B4	Bloods: urea	bun	2.5 $\leq$ bun $\leq$ 7.8 (mmol/L)

Table 1: Codified version of the discharge criteria for application to electronic health record data. Here the fifteen criteria have been grouped into intuitive subsets and each assigned a test ID ('R0' to 'B4'). If all 15 criteria are met for a period of at least four hours the patient can be safely discharged.

Cohort selection

Subjects for this study were selected from two distinct historical data sources to form two patient cohorts. The inclusion criteria are detailed in section B of the online supplementary file. The first data source consists of the routinely collected data from 1933 patients treated on the general intensive care unit at the Bristol Royal Infirmary between 01/02/2015 and 01/02/2017. We refer to the cohort selected from this dataset as GICU. The second data source is derived from the MIMIC-III database[12], from which 10658 patients were selected to form the cohort referred to as MIMIC.

The use of two cohorts was motivated by two concerns. Firstly, the volume of data was significantly increased by the inclusion of the MIMIC cohort, increasing the volume of data available for training classifier algorithms. Secondly, the use of two cohorts allowed us to study the generalisation of our results between different patient populations under different healthcare systems.

Readiness-for-discharge

The key to testing and improving on discharge criteria was to be able to identify, from the historical data, patients that were ready-for-discharge (RFD) and not-ready-for-discharge (NRFD). These two subsets of patients, RFD and NRFD, define the positive and negative classes respectively. The datasets contain a callout for each patient, which marks the time at which a patient was declared clinically ready to leave the intensive care unit. For the positive class we selected patients at their time of callout who went on to have a positive outcome on leaving the unit. Conversely, patients who were declared clinically ready to leave the unit, but subsequently had a negative outcome, were included in the negative class. A positive outcome was defined as the patient leaving the hospital alive without readmission to ICU. A negative outcome was defined as either readmission to ICU during the same hospital admission, or in-hospital mortality after discharge from ICU. Given the low rates of negative outcome following callout in both MIMIC and GICU (see table 2), we generated further instances of the negative class, in order to balance the class sizes. To do this we sampled patients at between three and eight days prior to their callout (see supplementary section B: figures 1-3), under the assumption that patients were not-ready-for-discharge at this point in time, regardless of their eventual outcome state (positive or negative). Full details of this procedure are given in section B of the online supplementary file.

	patients	mortality	readmission	negative outcomes	mean(LOS)
MIMIC	10658	0.048	0.063	0.095	3.02
GICU	1933	0.038	0.031	0.062	4.92

Table 2: Summary of the two study cohorts: MIMIC and GICU. Negative outcome after discharge is defined as either readmission to ICU or patient mortality during the same hospital admission (not mutually exclusive). Length of stay(LOS) given in days.

Feature extraction

We used the same feature set to evaluate the NLD criteria and to train machine learning classifiers. We constructed either one or two features corresponding to each of the NLD criteria, depending on the criteria in question and on data availability. For example, the features ‘resp min’ and ‘resp max’ were used to test the criterion R4, whereas the single feature ‘bun’ was used to test B4. Where possible the feature values were calculated from a four hour sample window, as specified by the original NLD criteria. In the cases where no data was available during the four hour window, an extended 36 hour window was used. This extended window was mainly relevant for infrequently measured laboratory test results (see table 1 in section C of the online supplementary file). Full details and justification of the feature extraction procedure are provided in section C of the online supplementary file, and the resulting 18 features are listed in the first column of table 3.

The results presented in the main text represent a complete case analysis, with all instances containing missing data entries removed. This removal reduced the sizes of the MIMIC and GICU feature matrices to 5033 and 1858 instances respectively. The validity of the complete case analysis was investigated with a sensitivity analysis that used k-nearest neighbour imputation[13] to fill missing data (see section D of the online supplementary file). When training and testing the machine learning classifiers, features were standardised by subtracting the column mean and dividing by the standard deviation. The complete case feature matrices are visualised in figure 1 using the t-SNE algorithm[14] (see supplementary section D: figure 4 for the equivalent imputed feature matrices).

Analysis of NLD criteria

Knight originally specified that all 15 criteria must be met in order to allow safe discharge by a nurse[7]. Following this specification we evaluated the criteria for both MIMIC and GICU, determining which instances were classified as RFD and NRFD, and comparing these results to ground-truth. We then further investigated the performance of the NLD criteria as a classification system, by relaxing the constraint that all 15 tests must be passed in order to make an RFD classification. Instead we used the NLD criteria to produce probability estimates of being RFD, by summing the number of tests passed and dividing by 15 to produce a normalised output between 0 and 1. In this formulation each of the 15 criteria contribute equally to the RFD probability. Using the probability outputs it was possible to evaluate the performance of the NLD criteria in the same way as the machine learning classifiers described below.

Machine learning classifiers

To improve upon the performance of the NLD criteria, we trained and tested two machine learning classifiers: a random forest (RF)[15], and a logistic classifier (LC)[16]. These two algorithms were chosen for their simplicity in implementation and ease of interpretation in their predictive output. Both classifiers were optimised over a range of hyper-parameter values using cross-validation (see section E of the online supplementary file). To do so, we produced an ensemble of 100 randomised train:test data splits for both MIMIC and GICU (70:30 and 67:33 respectively). We then optimised each classifier using cross-validation, with the larger subset of MIMIC as training data and the larger subset of GICU as validation data. Having obtained the optimised classifier for the given data split, the classifier was then refitted to the full training set (MIMIC + GICU). The remaining data (i.e. the smaller subsets of MIMIC and GICU) were then used to test the performance of the optimised classifier. This optimisation/testing procedure was repeated for all 100 data splits in the ensemble to produce estimates of the mean and standard deviation of the classifier performance. The methodology described was intended to produce classifiers that generalised well from MIMIC to GICU, and by extension to other patient populations.

Both the RF and LR classifiers contained mechanisms to promote model sparsity. For the RF this mechanism was a restriction in the hyper-parameters *tree_depth* and *max_features* (see section E of the online supplementary file), while for the LC this was ‘l1’ -regularisation. In both cases the intention was to learn which features were most and least predictive of readiness-for-discharge. For the RF, feature importance was given by the total reduction in Gini impurity provided by each feature. Intuitively this metric captures the reduction in uncertainty brought about a given feature, therefore giving a measure of how important that feature is, on average, in classification decisions. For the LC, the importance was given by the absolute value of the coefficient for each feature in the trained model, capturing the effect of a small change in a given feature value on the prediction probability with all other feature values held constant.

Classifier performance was evaluated across a range of prediction thresholds by producing receiver-operator-characteristic (ROC) and precision-recall (PRC) curves[17]. Given the need to minimise the false positive rate, while retaining high recall, we chose to use the partial area under the ROC curve (pAUC) as the overall performance metric[18]. The pAUC was evaluated up to a false positive rate of 0.3, using linear interpolation to approximate the true positive rate at this point on the ROC curve. Performance was evaluated in this way for the RF and LR classifiers, and for the original NLD criteria.

Results.

The original specification of the NLD criteria proved to be highly conservative as expected, producing low false positive and true positive rates for both cohorts (supplementary section D: tables 2-5). The true positive rates for MIMIC and GICU were 0.4% and 6.4% respectively. As such the NLD criteria were sufficiently insensitive as to call into question their usefulness in their current form.

By relaxing the constraint that all 15 tests must be passed, the NLD criteria were able to successfully identify more patients as RFD. This is illustrated in figure 2 for a single train:test data split, alongside the performance of the corresponding optimal random forest classifier (see also supplementary section D: figure 5). On this data split the NLD criteria obtained precisions of ~0.8 at a recall of 0.4 for both cohorts. The performance gain obtained by using a random forest was significant, with precisions of >0.8 at a recall of 0.7 for both cohorts. At a false positive rate of 3.6% the NLD criteria produced 66 true positives, while the random forest produced 113.

Feature	Importance (RF)	Importance (LC)	Rank (RF)	Rank (LC)
gcs_min	0.352 (±0.054)	1.127 (±0.050)	0	0
airway	0.302 (±0.048)	0.870 (±0.038)	1	1
bun	0.034 (±0.007)	0.437 (±0.047)	4	2
fio2	0.048 (±0.007)	0.317 (±0.024)	2	3
temp_max	0.024 (±0.008)	0.263 (±0.126)	7	4
haemoglobin	0.029 (±0.006)	0.256 (±0.029)	5	5
resp_max	0.017 (±0.004)	0.236 (±0.038)	10	6
resp_min	0.045 (±0.015)	0.233 (±0.053)	3	7
temp_min	0.014 (±0.004)	0.213 (±0.126)	12	8
hr_min	0.022 (±0.005)	0.181 (±0.044)	8	9
hr_max	0.025 (±0.005)	0.168 (±0.044)	6	10
spo2_min	0.012 (±0.004)	0.158 (±0.027)	14	11
na	0.009 (±0.003)	0.110 (±0.032)	16	12
bp_min	0.012 (±0.004)	0.059 (±0.027)	13	13
hco3	0.010 (±0.004)	0.051 (±0.028)	15	14
k	0.008 (±0.003)	0.041 (±0.022)	17	15
creatinine	0.021 (±0.005)	0.031 (±0.027)	9	16
pain	0.016 (±0.008)	0.021 (±0.018)	11	17

Table 3: Feature importances given by the random forest (RF) and logistic classifier (LC), evaluated over 100 train:test data splits. Importance values are given as: mean(±standard deviation). Features are ranked according to mean importance value, and the table is ordered according to the ranking given by the logistic classifier.

Broadly the two classifiers agreed as to which features were most predictive of readiness-for-discharge (see table 3). Eight of the logistic classifier's top ten important features were also ranked

in the top ten by the random forest, when averaged over the ensemble of 100 data splits. The Spearman's rank correlation coefficient between the feature rankings was 0.800 (p=0.00006), and both classifiers ranked *gcs_min* and *airway* as the two most important features by a significant margin. The inclusion of instances with missing data did little to change these feature rankings (see supplementary section D: figures 6-7).

Figure 3 summarises classifier performance, quantified using the partial area under the ROC curve (pAUC), over the ensemble of data splits. On average the original NLD criteria performed slightly better for GICU than MIMIC. The two machine learning classifiers performed similarly well, producing large gains in pAUC over the NLD for both cohorts, and higher pAUC values for MIMIC than GICU. There was little to distinguish between the random forest and logistic classifiers based on pAUC. In the sensitivity analysis (see supplementary section D: figure 8 and table 6) both machine learning classifiers still performed better than the NLD criteria, but all classifiers showed a performance drop for MIMIC and one classifier tended to perform better for each cohort (the random forest and logistic classifiers for MIMIC and GICU respectively).

Given the similarity in classifier performances we chose to use the average feature importances of the simpler model – the logistic classifier - to weight the NLD criteria. The weighted version of the NLD criteria (referred to as NLD_{opt}) performed better than the original criteria when tested on both MIMIC and GICU. On MIMIC the performance gain was larger, with pAUC scores approaching those of the machine learning classifiers. Qualitatively the same effect was observed under the sensitivity analysis.

Examples in practice

To illustrate the results in a more human-interpretable fashion we have selected five informative examples from the GICU cohort. Table 4 summarises the performance of the different classification systems for these five examples, which are labelled as true or false positive/negative (TP,FP,TN,FN) according to how they would be classified under the original nurse-led discharge criteria. One patient (ID 4065) is included twice: once at 72 hours before callout, and again at the time of callout. All four classification systems show an increased RFD probability for this patient between the two time points, as would be expected. Patient 868 is a false negative under the original criteria - despite failing two criteria (C0 and T) their callout was successful. The three alternative classification systems (NLD_{opt} , RF and LC) correctly assign a high RFD probability for this patient, therefore improving upon the original criteria.

For these select examples the logistic classifier is the only system to assign lower RFD probabilities to the two false positive instances (1034 and 10783) than to the true positive instance (4065). Despite this correct ordering by the logistic classifier its RFD probabilities for the false positives are relatively high. Given the conservative nature of the NLD criteria it is expected that correct classification of the false positive instances is a hard problem.

Patient	NLD	NLD_{opt}	RF	LC	NLD fails	Notes
1034 (FP)	0.010 (1.0)	0.010 (1.0)	0.071 (0.784)	0.096 (0.765)	-	Patient admitted to ICU post surgery (primary lung tumour). Discharge to ward. Readmitted within 24 hours with bacterial pneumonia.
10783 (FP)	0.010 (1.0)	0.010 (1.0)	0.035 (0.819)	0.163 (0.716)	-	Patient admitted to ICU with secondary hepatic tumour. Appears to be RFD at 96 hours prior to callout.
4065 (TN)	1.0 (0.467)	0.464 (0.702)	0.368 (0.494)	0.395 (0.450)	R2, R4, C0, P, B1, B3, B4	Patient admitted to ICU with intracranial abscess. Not ready for discharge at 72 hours prior to callout.
4065 (TP)	0.010 (1.0)	0.010 (1.0)	0.077 (0.780)	0.047 (0.812)	-	Same patient as above. RFD at time of callout.
868 (FN)	0.113 (0.867)	0.046 (0.939)	0.080 (0.777)	0.054 (0.806)	C0, T	Patient admitted with malignant large bowel tumour. Appears NRFD at time of callout. Positive outcome.

Table 4: Example patients and their scores given by the four classification systems: the original nurse led discharge criteria (NLD); the weighted criteria(NLD_{opt}); the random forest (RF); and the logistic classifier (LC). The reporting score given is the false positive rate at the point where the patient falls on the ROC curve, such that a lower score indicates a higher probability of being RFD according to the given classifier (explicitly, the reporting metric gives the number of false positives that must be accepted before this patient can be classified RFD). These scores can be compared across classifiers. The raw (non-calibrated) classifier scores are given in brackets and cannot be compared across classifiers. The results FP,TN,TP, and FN indicated in the first column correspond to the outcomes of the original NLD criteria. The column 'NLD fails' specifies, where relevant, which of the NLD criteria were not met (criteria IDs correspond to those in table 1).

Discussion

Identifying which patients are suitable for ICU discharge is complex[1]. Delayed and out of hours discharges are associated with an increased mortality[19], and patients in ICU who could be managed on the ward put an increasing strain on resources. The determination of ready-for-discharge status is influenced by many unmeasured factors such as ICU census[20] and this leads to unwarranted variation in clinical decision making. The decision to declare someone fit for discharge is based on the judgement of individual clinicians and is likely to be given a lower priority than decisions around treatment options for patients that are more unwell in the ICU.

In this study we have used routinely collected data to test a set of discharge criteria[7] against two machine learning classifiers. The discharge criteria were found to be highly conservative, with very few successfully discharged patients meeting all the required criteria. This low sensitivity was expected since the original criteria were designed to be implemented independently of usual ward rounds, and false positives in this scenario could have serious consequences. A random forest and a logistic classifier both performed better than the clinically derived discharge criteria when trained on the same feature set. The two classifiers broadly agreed on which features were most predictive of readiness-for-discharge. Weighting the original discharge criteria with the features importances of the logistic classifier improved their classification performance.

An important novel aspect of this work is the use of MIMIC-III to increase the volume of training data available locally. Such applications of machine learning techniques to datasets that span institutions and healthcare settings will be of increasing value as more intensive care datasets become available for research[21]. Our results demonstrate the feasibility of using combined datasets in this way to derive clinical insight, and could be developed by the application of transfer learning approaches[22] to characterise systematic differences between data distributions.

The features identified as important by the classifiers were clinically meaningful. Clinicians will recognise that coma score, respiratory function and renal function are strongly related to successful ICU discharge. It is perhaps surprising that cardiovascular parameters were not ranked higher. We propose two possible mechanisms to account for this apparent discrepancy. Firstly, it may be a consequence of patient heterogeneity on the general intensive care unit[23]. For example, cardiovascular parameters may be highly predictive for cardiac patients yet much of this predictive power is lost in our attempt to fit a general model for the whole ICU population. Secondly, it may be a due to our simplistic choice of features, which use the absolute values of physiological parameters. For some parameters we suggest that other features such as the trend, variance, or change since time of admission may be more predictive. For example, improvement in blood pressure may be more informative than absolute blood pressure.

Our feature set was chosen to be directly analogous with those used by Knight's criteria, to allow a direct comparison in performance. This feature set is somewhat restrictive, having been originally designed to be manually recorded by nurses using paper charts. The rich wealth of data held in electronic charting systems could be better exploited by including more physiological parameters, and engineering more predictive features. In particular our modelling did not make use of demographic information, diagnoses, comorbidities or interventions. The later is of particular importance since many of patient's physiological parameters are controlled by clinical intervention during their stay in ICU. For example, a patient on vassopressors may have close to normal blood pressure despite suffering form server cardiovascular complications. Therefore, conditioning features on medical interventions represents one avenue to significantly boost performance. Methods to account for patient heterogeneity and individual disease trajectories would also be worth investigating[23,24]. Although the inclusion of entries with missing data did not qualitatively alter the results of our complete case analysis, the development of a robust imputation

strategy would improve performance by making best use of the available training data and exploiting the value in missingness[25].

The aggregate effects of the improvement gains produced by our machine learning approach could be beneficial to many[26]. Therefore, we suggest that a future decision support tool embedded within a CIS should use these techniques to alert clinicians when patients appear fit for discharge. The increasing worldwide adoption of information systems in intensive care would make such a system widely applicable in years to come[27]. We have shown in previous work that subtle changes to the presentation of information can have significant impact on clinical decision making[28]. Therefore we anticipate such a tool has the potential to significantly streamline the discharge process. Two issues would need to be addressed prior to implementation on the ICU. The first is the human-interpretability of the classifier output. Depending on the machine learning approach a number of solutions exist[29], including the approximation of random forests with simple decision trees[30], that would allow clinicians to engage with the reasons behind a given classification. The second is the ambiguity behind the ground-truth used to train the classifiers. For example, some patients in the datasets used did not have a recorded callout status and some patients may have been ready for discharge prior to callout. A live implementation with a human-in-the-loop[31] could use clinician input to update ground-truth in such situations and improve learning.

Conclusion

We have shown that it is possible to apply machine learning techniques to routinely collected ICU data in order to solve a significant clinical and operational problem. This approach offers promise in a number of areas. We plan to focus on the development and deployment of a decision support tool in order to inform clinicians of patients that could potentially be discharged from ICU, in order to streamline the process and reduce unnecessary ICU stay. As more patient data becomes available in the wider hospital setting, there is extensive scope to use such data-driven methods to solve the problem of poor patient flow through hospitals.

References

1 Rubenfeld GD, Rhodes A. How many intensive care beds are enough? *Intensive Care Med* 2014;**40**:451–2. doi:10.1007/s00134-014-3215-x

2 Chalfin DB, Trzeciak S, Likourezos A, *et al*. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Crit Care Med* Published Online First: 2007. doi:10.1097/01.CCM.0000266585.74905.5A

3 Cardoso LT, Grion CM, Matsuo T, *et al*. Impact of delayed admission to intensive care units on mortality of critically ill patients: a cohort study. *Crit Care* 2011;**15**:R28. doi:10.1186/cc9975

4 Howell MD. Managing ICU throughput and understanding ICU census. *Curr Opin Crit Care* 2011;**17**:626–33. doi:10.1097/MCC.0b013e32834b3e6e

5 Capuzzo M, Moreno RP, Alvisi R. Admission and discharge of critically ill patients. *Curr Opin Crit Care* 2010;**16**:499–504. doi:10.1097/MCC.0b013e32833cb874

6 Stelfox HT, Lane D, Boyd JM, *et al*. A Scoping Review of Patient Discharge From Intensive Care. *Chest* 2015;**147**:317–27. doi:10.1378/chest.13-2965

7 Knight G. Nurse-led discharge from high dependency unit. *Nurs Crit Care*;**8**:56–61.

8 Obermeyer Z, Lee TH. Lost in Thought — The Limits of the Human Mind and the Future of Medicine. *N Engl J Med* 2017;**377**:1209–11. doi:10.1056/NEJMp1705348

9 Docherty AB, Lone NI. Exploiting big data for critical care research. *Curr Opin Crit Care* 2015;**21**:467–72. doi:10.1097/MCC.0000000000000228

10 Sohn E, Roski J, Escaravage S, *et al*. Four Lessons In The Adoption Of Machine Learning In Health Care. *Heal Aff Blog* 2017. doi:10.1377/hblog20170509.059985

11 Kawamoto K, Houlihan CA, Balas EA, *et al*. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ* 2005;**330**:765. doi:10.1136/bmj.38398.500764.8F

12 Johnson AEW, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;**3**:160035. doi:10.1038/sdata.2016.35

13 Batista GEAPA, Monard MC. An analysis of four missing data treatment methods for supervised learning. *Appl Artif Intell* 2003;**17**:519–33. doi:10.1080/713827181

14 Maaten L van der, Hinton G. Visualizing Data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.

15 Liaw A, Wiener M. Classification and Regression by RandomForest. *R News* 2002;**2**:18–22.

16 Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform*;**35**:352–9.

17 Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd international conference on Machine learning - ICML '06*. New York, New York, USA: : ACM Press 2006. 233–40. doi:10.1145/1143844.1143874

18 Zou KH, O'Malley AJ, Mauri L. Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. *Circulation* 2007;**115**:654–7. doi:10.1161/CIRCULATIONAHA.105.594929

19 Goldfrad C, Rowan K. Consequences of discharges from intensive care at night. *Lancet (London, England)* 2000;**355**:1138–42. doi:10.1016/S0140-6736(00)02062-6

20 Lin F, Chaboyer W, Wallis M. A literature review of organisational, individual and teamwork factors contributing to the ICU discharge process. *Aust Crit Care* 2009;**22**:29–43. doi:10.1016/j.aucc.2008.11.001

21 Harris S, Shi S, Brealey D, *et al.* Critical Care Health Informatics Collaborative (CCHIC): Data, tools and methods for reproducible research: A multi-centre UK intensive care database. *Int J Med Inform* 2018;**112**:82–9. doi:10.1016/j.ijmedinf.2018.01.006

22 Pan SJ, Yang Q. A Survey on Transfer Learning. *IEEE Trans Knowl Data Eng* 2010;**22**:1345–59. doi:10.1109/TKDE.2009.191

23 Vranas KC, Jopling JK, Sweeney TE, *et al.* Identifying Distinct Subgroups of ICU Patients. *Crit Care Med* 2017;**45**:1607–15. doi:10.1097/CCM.0000000000002548

24 Alaa AM, Yoon J, Hu S, *et al.* Personalized Risk Scoring for Critical Care Prognosis Using Mixtures of Gaussian Processes. *IEEE Trans Biomed Eng* 2018;**65**:207–18. doi:10.1109/TBME.2017.2698602

25 Lin J-H, Haug PJ. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. *J Biomed Inform* 2008;**41**:1–14. doi:10.1016/J.JBI.2007.06.001

26 Parker M. The Aggregation of Marginal Gains. *Bull R Coll Surg Engl* 2011;**93**:236–7. doi:10.1308/147363511X582239

27 Mador RL, Shaw NT. The impact of a Critical Care Information System (CCIS) on time spent charting and in direct patient care by staff in the ICU: A review of the literature. *Int J Med Inform* 2009;**78**:435–45. doi:10.1016/j.ijmedinf.2009.01.002

28 Bourdeaux CP, Thomas MJ, Gould TH, *et al.* Increasing compliance with low tidal volume ventilation in the ICU with two nudge-based interventions: evaluation through intervention time-series analyses. *BMJ Open* 2016;**6**:e010129. doi:10.1136/bmjopen-2015-010129

29 Ribeiro MT, Singh S, Guestrin C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. New York, New York, USA: : ACM Press 2016. 1135–44. doi:10.1145/2939672.2939778

30 Deng H. Interpreting tree ensembles with inTrees. *Int J Data Sci Anal* 2018;:1–11. doi:10.1007/s41060-018-0144-8

31 Holzinger A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 2016;**3**:119–31. doi:10.1007/s40708-016-0042-6

Figure captions:

Figure 1: A single t-SNE embedding of the two feature matrices with each cohort plotted separately: GICU (left), and MIMIC (right). Green and red points indicate instances of RFD and NRFD respectively. The more similar two instances (in terms of feature values), the closer together they appear in the embedding space. Feature matrices displayed are those with missing entries removed.

Figure 2: Performance of the random forest (RF) and NLD classifiers over a range of prediction thresholds, for a single train:test data split. RF optimised using cross-validation with data from MIMIC and GICU (see main text). Performance evaluated on the test subset from MIMIC and GICU.

Figure 3: Performance of the four classification systems. Performance metric is the partial area under the ROC curve up to a false positive rate of 0.3. Orange lines show median pAUC value over the ensemble, while the boxes indicate the lower and upper quartiles.

Contributors:

CJM is the main author and conducted the data processing and analysis. RSR and DJL made important technical and methodological contributions. IDG, AC, CPB drove the study concept and design. The clinical expertise of THG, MJCT and CPB informed all stages of the project, in particular study design and interpretation of results. All authors contributed to writing the manuscript and approved the final version.

Funding:

CJM was funded by the Elizabeth Blackwell Institute Catalyst Fund. DJL is funded by Wellcome Trust and Royal Society Grant Number WT104125MA.

Acknowledgements:

We would like to thank Graeme Palmer and Amy Weaver for their support in accessing and understanding the GICU data.

Competing interests:

None declared.

Ethics approval:

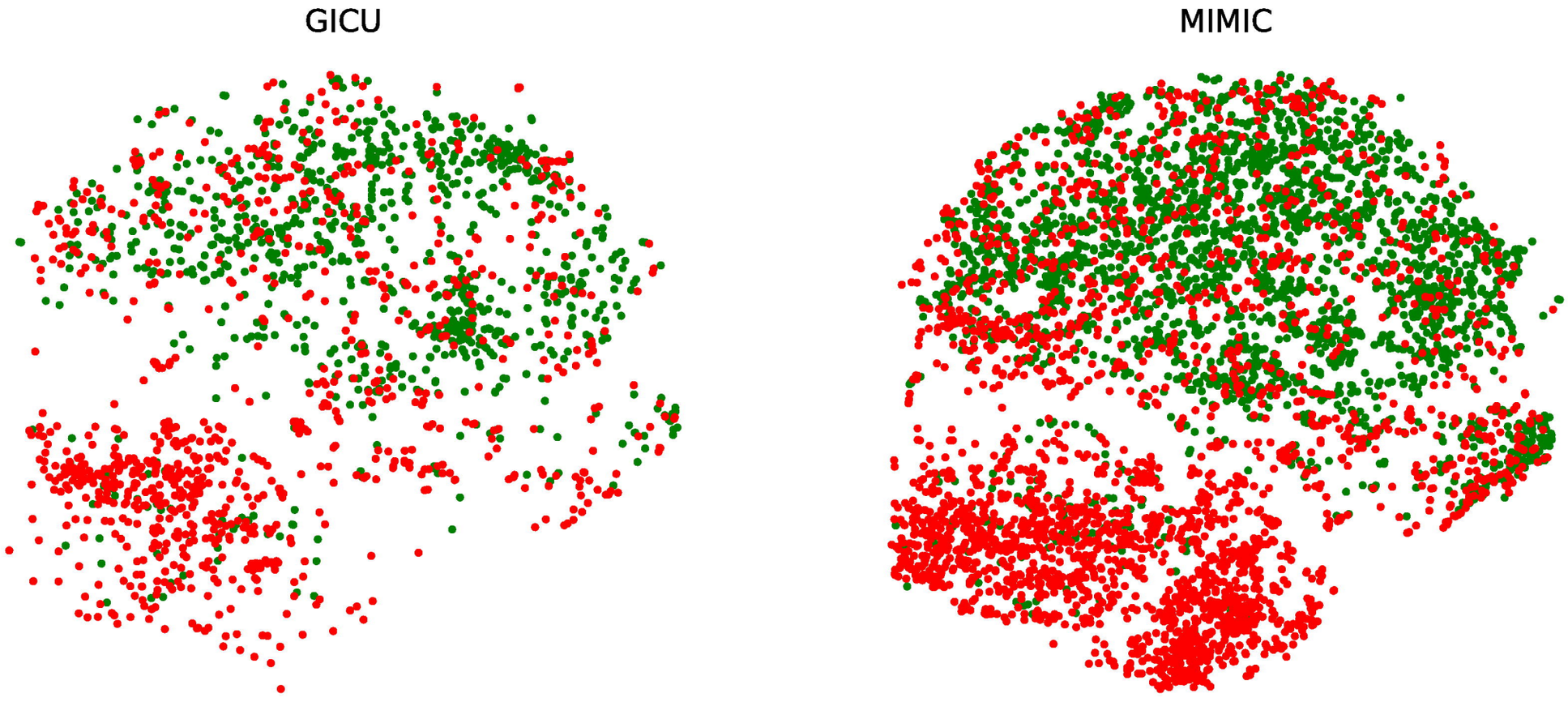
University Hospitals Bristol NHS Foundation Trust audit and service evaluation.

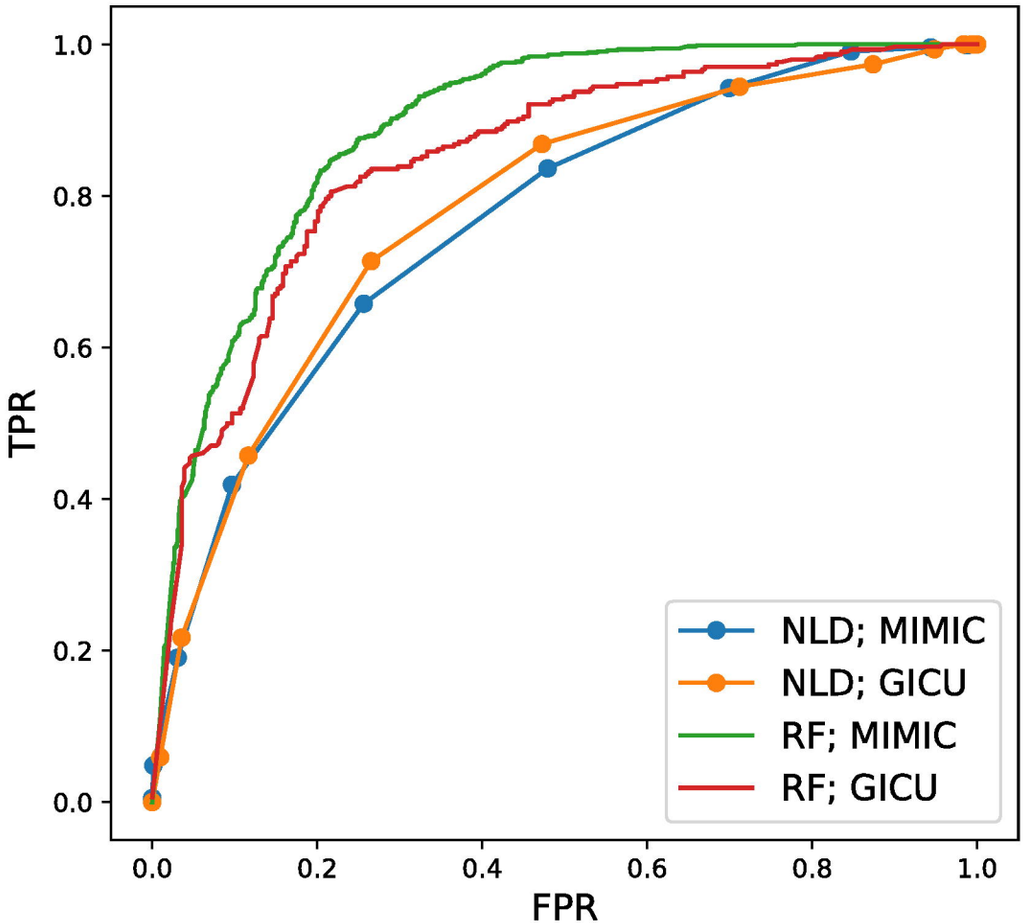
Patient and Public Involvement :

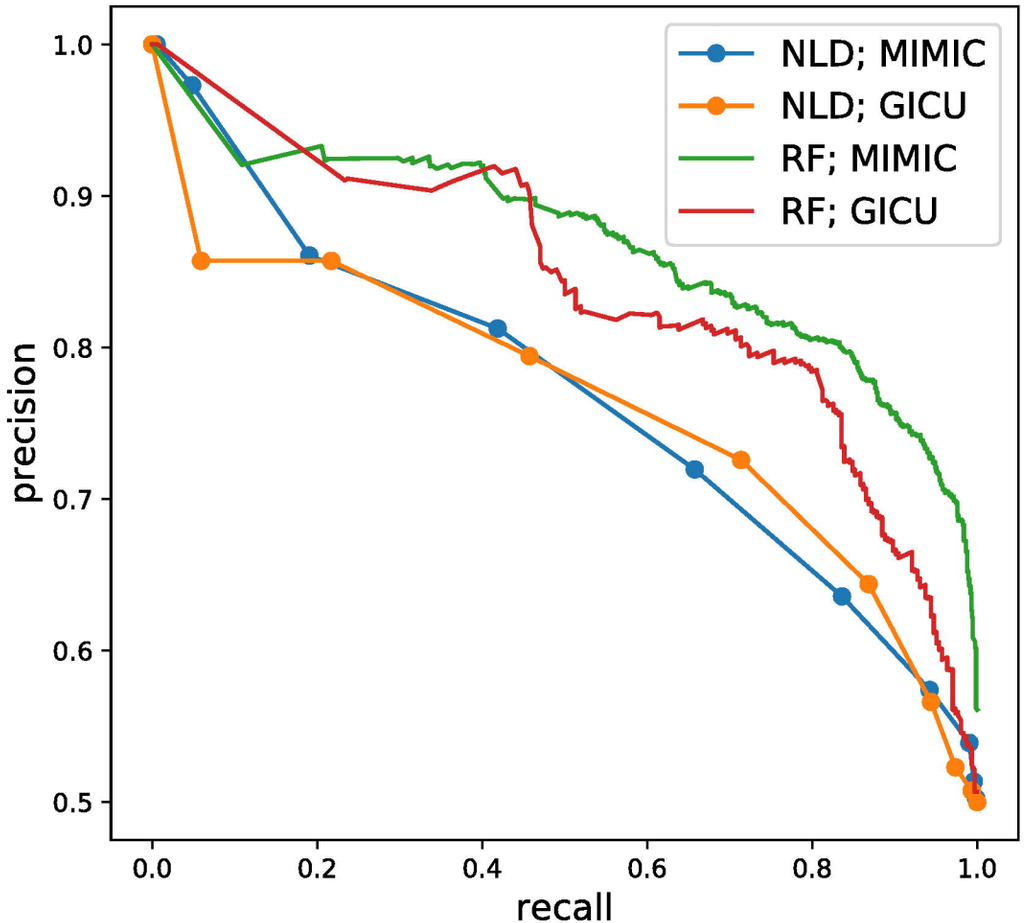
Patients were not directly involved in this study and patient consent was not required.

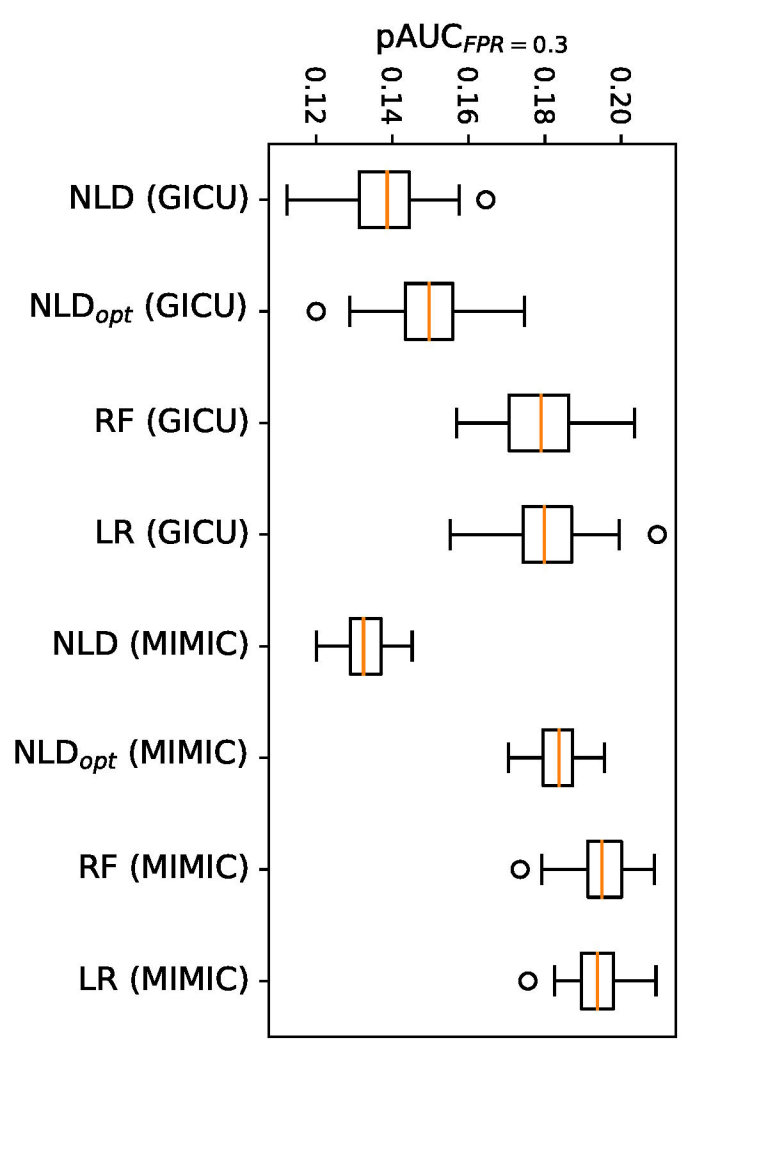
Data sharing statement:

Feature matrices will be made available on Dryad. Python code for analysis and processing on request directly from the corresponding author.


GICU
MIMIC


TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
NLD; MIMIC
NLD; GICU
RF; MIMIC
RF; GICU


precision
recall
0.5
0.6
0.7
0.8
0.9
1.0
NLD; MIMIC
NLD; GICU
RF; MIMIC
RF; GICU


pAUC$_{FPR = 0.3}$
0.12
0.14
0.16
0.18
0.20
NLD (GICU)
NLD$_{opt}$ (GICU)
RF (GICU)
LR (GICU)
NLD (MIMIC)
NLD$_{opt}$ (MIMIC)
RF (MIMIC)
LR (MIMIC)