**A machine learning approach to intensive care discharge.**

Christopher J. McWilliams∗1, Daniel J. Lawson2, Raul Santos-Rodriguez1, Iain D Gilchrist, Alan Champneys1, Timothy H. Gould4, Matthew J.C. Thomas4, and Christopher P. Bourdeaux4

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](mailto:chris.mcwilliams@bristol.ac.uk)

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).

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 the resource is constrained[1]⁠. In order to optimise the allocation of this resource, patients should be discharged from the ICU 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 patients requiring critical care[2,3]⁠⁠. Furthermore, patients remaining in the ICU after they are medically fit to leave are at risk of iatrogenic harm and may 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]⁠. A number of studies have looked to address this problem by attempting to standardise the discharge process

In a scoping review of these studies Stelfox et al.[6]⁠ noted that, while a range of tools have been developed to characterise discharge readiness, most studies have been single centre and have not conducted comparative evaluations of different tools.

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[7,8]⁠. Two particular avenues for development are the automation of simple tasks[9]⁠ and the implementation of decision support systems[10]⁠, both of which could 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 improved by such data driven approaches. Indeed, several statistical models have recently been developed to predict the risk of adverse events following intensive care discharge[11–15]⁠. Such risk models are invaluable tools for clinical decision making, and in the context of ICU discharge can provide information with which to plan complex de-escalations of care. For example, patients deemed to be at high-risk of readmission may benefit from continued close monitoring[16]⁠, since early detection of deterioration is a strong predictor of outcome[17,18].

In our previous work on the psychology of clinical decision making we have demonstrated the effectiveness of simple ‘nudge’ based interventions in changing clinical practice[19–21]⁠⁠. Building on this foundation we were motivated to develop a classifier to automatically flag patients that appear physiologically fit for discharge. The intention is that such a screening tool could expedite morning ward rounds by allowing staff to focus their attention on the most likely-dischargeable patients. The tool could also prompt clinicians to consider discharge decisions at other times of day, outside of normal rounds. In 2003 Knight proposed a set of nurse-led discharge criteria[22]⁠ with a similar aim – to expedite discharge from a high-dependency unit by allowing nurses to discharge patients who were clearly well enough to leave. These criteria 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). High-risk patients are unlikely to meet these criteria, but may still be discharegable by a consultant. In this study we used routinely collected patient data to retrospectively evaluate Knight’s criteria, and then improve upon their performance using machine learning methods. To this end 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[23]⁠ (see Materials and Methods for details).

Methods

Discharge criteria

The nurse-led discharge (NLD) criteria proposed by Knight[22]⁠ consist of a set of constraints on various routinely collected vital signs and laboratory results. If a patient meets all the criteria for a period of at least four hours, Knight states that they may be safely discharged by a nurse. The motivation behind developing these criteria was to facilitate discharge by nurses in cases where the decision was clear, and there is some evidence of improved bed allocation when using such a nurse-led system[22,24,25]⁠⁠. 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. For criteria that were not assigned numeric values in the original publication (B1-4, CNS) we used the ‘normal’ bounds as defined in our clinical information system. Limitations and potential improvements to these criteria are addressed in the discussion.

|  |  |  |  |
| --- | --- | --- | --- |
| Test ID | Test name | Variable | Test condition |
| R0 | Respiratory: airway | airway | airway patent |
| R1 | Respiratory: FiO2 | fio2 | fio2 ≤ 0.6 |
| R2 | Respiratory: blood oxygen | spo2 | spo2 ≥ 95 (%) |
| R3 | Respiratory: bicarbonate | hco3 | hco3 ≥ 19 (mmol/L) |
| R4 | Respiratory: rate | resp (rate) | 10 ≤ resp ≤ 30 (bpm) |
| C0 | Cardiovascular: blood pressure | bp (systolic) | bp ≥ 100 (mmHg) |
| C1 | Cardiovascular: heart rate | hr | 60 ≤ hr ≤ 100 (bpm) |
| P | Pain | pain | 0 ≤ pain ≤ 1 |
| CNS | Central nervous system | gcs | gcs ≥ 14 |
| T | Temperature | temp | 36 ≤ temp ≤ 37.5 (C) |
| B0 | Bloods: haemoglobin | haemoglobin | haemoglobin ≥ 9 (g/dL) |
| B1 | Bloods: potassium | k | 3.5 ≤ k ≤ 6.0 (mmol/L) |
| B2 | Bloods: sodium | na | 130 ≤ na ≤ 150 (mmol/L) |
| B3 | Bloods: creatinine | creatinine | 59 ≤ creatinine ≤ 104 (umol/L) |
| B4 | Bloods: urea | bun | 2.5 ≤ bun ≤ 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 was derived from the MIMIC-III database[23]⁠, from which we selected patients who were admitted to medical or surgical intensive care since this approximates the patient type in GICU. We restricted our analysis to the ‘Metavision’ subset of MIMIC-III, since the labelling of the variables required to evaluate the NLD criteria was found to be more consistent in this portion of the database. Furthermore, we selected only the first intensive care stay of any given hospital admission, and only those stays for which there was a recorded *callout* (ready-for-discharge) time. Following these criteria we arrived at a subset of 10658 patients from MIMIC-III, forming the cohort we refer to hereafter as *MIMIC*.

The use of two cohorts was motivated by two concerns. Firstly, by including the MIMIC cohort, we significantly increased 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 the 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). Whereas previous models have looked to predict the occurrence of adverse events following ICU discharge[12,15]⁠ we wanted to learn to classify those patients that appear physiologically fit for discharge. These are subtly different tasks. The former requires the identification of patients at risk of negative outcomes from those who have already been declared fit for discharge, while the later looks to identify, from a sample of ICU patients, those patients that are no longer in need of critical care. Clearly the later is an easier task. In order to train a classifier for this task it was necessary to define instances of the positive (RFD) and negative classes (NRFD). Both datasets (GICU and MIMIC) contain a callout for each patient, which marks the time at which a patient was declared clinically ready to leave the intensive care unit. A patient was defined as RFD at their time of callout, provided they had a positive outcome after leaving ICU. Conversely, patients with a negative outcome were defined as NRFD at their time of callout. A positive outcome was defined as the patient leaving 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. Potential issues with these definitions are addressed in the discussion.

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. Conceptually this is equivalent to providing more instances for the classifier to learn the physiological characteristics of patients requiring ongoing critical care. To do this we sampled patients at between three and eight days prior to their callout, under the assumption that patients were not-ready-for-discharge at this point in time, regardless of their eventual outcome state (positive or negative). Patients within the first 24 hours of their ICU stay were omitted from this sample. Full details of the sampling 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.

To produce the results presented in the main text, missing feature values were imputed using k-nearest neighbour imputation[26]⁠. Full details of the imputation procedure are given in section D of the online supplementary file, along with a complete case analysis that addresses the sensitivity of our results to the missing values. When training and testing the machine learning classifiers, features were standardised by subtracting the mean and dividing by the standard deviation. The feature matrices are visualised in figure 1 using the t-SNE algorithm[27]⁠.

Analysis of NLD criteria

Knight originally specified that all 15 criteria must be met in order to allow safe discharge by a nurse[22]⁠. 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. Subsequently we weighted each criteria according to a measure of feature importance (see below) in order to improve their predictive performance. 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)[28]⁠, and a logistic classifier (LC)[29]⁠. These two algorithms were chosen for their simplicity in implementation and ease of interpretation in their predictive output. The training methodology we used was intended to produce classifiers that made the good use of the training data, derived from multiple source domains, whilst generalising well to new patient populations. As such we employed multiple-source cross-validation[30]⁠. A single iteration of this procedure is as follows. Each source dataset is split into train and test data. For GICU 30% of the data is held out for testing. For MIMIC an equal sized test set is held out (~10%). Multiple-source cross-validation is then used to optimise the hyper-parameters on the training data (see section E of the online supplementary file) with two validation folds, one derived entirely from MIMIC and the other derived entirely from GICU. The optimised classifier is then retrained on the full training data (MIMIC and GICU), and its performance is tested on the held out test data. This procedure is repeated over 100 random train-test splits to produce estimates of the mean and standard deviation of classifier performances.

In order to determine the feature importances for each classifier, and therefore understand which features were most predictive of readiness-for-discharge, we used the permutation feature importance[31]⁠. In short this procedure involves iterative random permutation of the values of each feature, and the calculation of average loss of classifier performance (we used area under the ROC curve) resulting from this feature randomisation. The overall performance of our trained classifiers, and the NLD criteria, was characterised by producing receiver-operator-characteristic (ROC) and precision-recall (PRC) curves[32]⁠, and by evaluating a suite of common performance metrics.

**Put this somewhere else:**

Classifier performance was evaluated across a range of prediction thresholds by producing . 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[33]⁠. 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.

Table 3: Patient characteristics for the cohorts used to produce the results in the main text. Discharge delay defined as length of time between callout and discharge from ICU. Readmission to ICU defined as readmission during same hospital stay. Negative outcome is mortality and/or readmission.  






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 (see confusion matrices in table 1 of appendix A). 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. 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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 4: 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 (appendixF).

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 figure 2 and table 1 in appendix F)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 NLDopt) 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 (NLDopt , 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 | NLDopt | 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 5: Example patients and their scores given by the four classification systems: the original nurse led discharge criteria (NLD); the weighted criteria(NLDopt); 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[34]⁠, and patients in ICU who could be managed on the ward put an increasing strain on resources. The determination of the ready-for-discharge status is influenced by many unmeasured factors such as ICU census[25]⁠ 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[22]⁠ 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[35]. 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[36]⁠ 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[37]⁠. 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[37,38]⁠⁠. 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[39]⁠.

The machine learning approach produced small absolute performance gains over the original discharge criteria. However, the aggregate effect of such small improvements could be beneficial to many[40]⁠. Therefore we suggest that a future decision support tool embedded within a CIS should use machine learning 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[41]⁠. We have shown in previous work that subtle changes to the presentation of information can have significant impact on clinical decision making[20]⁠. 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[42]⁠, including the approximation of random forests with simple decision trees[43]⁠, 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[44]⁠ 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 these methods to solve the problem of poor patient flow through hospitals.

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 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

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

9 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

10 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

11 Rojas JC, Carey KA, Edelson DP, *et al.* Predicting intensive care unit readmission with machine learning using electronic health record data. *Ann Am Thorac Soc* Published Online First: 2018. doi:10.1513/AnnalsATS.201710-787OC

12 Badawi O, Breslow MJ. Readmissions and Death after ICU Discharge: Development and Validation of Two Predictive Models. *PLoS One* Published Online First: 2012. doi:10.1371/journal.pone.0048758

13 Badawi O, Liu X, Hassan E, *et al.* Evaluation of ICU risk models adapted for use as continuous markers of severity of illness throughout the ICU stay. *Crit Care Med* Published Online First: 2018. doi:10.1097/CCM.0000000000002904

14 Cosgriff CV, Celi LA, Sauer CM. Boosting Clinical Decision-making: Machine Learning for Intensive Care Unit Discharge. *Ann Am Thorac Soc* 2018;**15**:804–5. doi:10.1513/AnnalsATS.201803-205ED

15 Desautels T, Das R, Calvert J, *et al.* Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: A cross-sectional machine learning approach. *BMJ Open* Published Online First: 2017. doi:10.1136/bmjopen-2017-017199

16 Jeffs E, Vollam S, Young JD, *et al.* Wearable monitors for patients following discharge from an intensive care unit: practical lessons learnt from an observational study. *J Adv Nurs* 2016;**72**:1851–62. doi:10.1111/jan.12959

17 Escobar GJ, LaGuardia JC, Turk BJ, *et al.* Early detection of impending physiologic deterioration among patients who are not in intensive care: Development of predictive models using data from an automated electronic medical record. *J Hosp Med* 2012;**7**:388–95. doi:10.1002/jhm.1929

18 Harris S, Singer M, Rowan K, *et al.* Delay to admission to critical care and mortality among deteriorating ward patients in UK hospitals: a multicentre, prospective, observational cohort study. *Lancet* 2015;**385**:S40. doi:10.1016/S0140-6736(15)60355-5

19 Bourdeaux CP, Davies KJ, Thomas MJC, *et al.* Using ‘nudge’ principles for order set design: a before and after evaluation of an electronic prescribing template in critical care. *BMJ Qual Saf* 2014;**23**:382–8. doi:10.1136/bmjqs-2013-002395

20 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

21 Bourdeaux CP, Birnie K, Trickey A, *et al.* Evaluation of an intervention to reduce tidal volumes in ventilated ICU patients † ‡. *Br J Anaesth* 2015;**115**:244–51. doi:10.1093/bja/aev110

22 Knight G. Nurse-led discharge from high dependency unit. *Nurs Crit Care* 2003;**8**:56–61.

23 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

24 Crocker C, Keller R. Nurse-led discharge to the ward from high dependency: A service improvement project. *Intensive Crit Care Nurs* 2005;**21**:363–6. doi:10.1016/J.ICCN.2005.05.004

25 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

26 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

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

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

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

30 Geras KJ, Sutton C. Multiple-source cross-validation. 2013.

31 Breiman L. Random Forests. *Mach Learn* 2001;**45**:5–32. doi:10.1023/A:1010933404324

32 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

33 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

34 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

35 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

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

37 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

38 Alaa AM, Yoon J, Hu S, *et al.* Personalized Risk Scoring for Critical Care Patients using Mixtures of Gaussian Process Experts. 2016.

39 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

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

41 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

42 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

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

44 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