Supplemental Digital Content 1

Explainable machine learning on AmsterdamUMCdb for ICU discharge decision support: uniting intensivists and data scientists

Patrick J. Thoral, MD[1], Mattia Fornasa, PhD[2], Daan P. de Bruin, MSc[2], Michele Tonutti, MRes[2], Hidde Hovenkamp, MSc[2], Ronald H. Driessen[1], Armand R.J. Girbes, MD, PhD[1], Mark Hoogendoorn, PhD[3], Paul W.G. Elbers, MD, PhD[1]

Affiliations

[1] Department of Intensive Care Medicine, Laboratory for Critical Care Computational Intelligence (LCCCI), Amsterdam Medical Data Science (AMDS), Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands

[2] Pacmed BV, Plantage Middenlaan 62, 1018 DH Amsterdam

[3] Computational Intelligence Group, Department of Computer Science, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081 HV, Amsterdam, The Netherlands

Contents

Feature engineering

Since there is a large difference in sample frequency of the data, ranging from every minute for life support devices to scores that are performed only once a week if ever, we specifically chose aggregation functions that could capture trends, extremes and availability for all relevant features. The choice of feature aggregates was inspired by previous work [1–3]. Each variable was binned in windows of 24 hours, starting from the moment of admission. For low-frequency laboratory tests, we imputed the values of the last window that was not empty. Then, three sets of features were computed in order to capture different aspects: (1) Short-term (day-level) data, which describe the condition of the patient on a specific day. Nine features were computed for each time window. We retain the features for the first and the last day, since all patients are expected to have these data. The former is expected to describe the patient's severity of illness at the moment of admission and the latter the condition at possible discharge; (2) Long-term (admission-level) data, included to detect trends over the complete ICU stay. Twelve long-term features were computed from the day-level aggregates and the difference between the day-level aggregations of the first and last day was calculated for all features as well; (3) Deviations from population-averaged patients, under the hypothesis that uncommon values (with respect to the distribution in the population) could be associated with readmission or mortality. For each 24-hour window, the patient's average was subtracted from the population's average for that specific 24-hour window. Using these values, we computed the same nine features as the short-term features (see Table S1).

Furthermore, medical expertise was used to identify features that showed a clear non-linear (U-shaped) relationship to the outcome on visual inspection of scatter plots (see Table S2). For those variables, additional features were created by calculating the squares. Finally, for each medication and supportive care ('treatment'), four features were computed: the total amount of time the patient received the treatment, the time between the last use of the treatment and discharge in both hours and as a fraction of the length of stay and a binary indicator to describe whether the patient was still receiving the treatment at ICU discharge. Missing values were imputed using mean imputation. Additionally, we created features indicating whether or not a specific variable was measured and features indicating how often this variable was measured to allow *informative missingness* to be modelled. Features were scaled to a distribution with zero mean and standard deviations equal to 1 before training the models.

Aggregations		Function
Short-term	**Day-level**	Mean
		Minimum
		Maximum
		Standard deviation
		First
		Last
		Count (number of measurements)
		Binary (yes/no)
		Slope
Long-term	**Admission-level**	Mean
		Minimum
		Maximum
		Count (number of measurements)
		Standard deviation
		Binary (yes/no)
	Average of day-level aggregations	Mean
		Minimum
		Maximum
		Count (number of measurements)
		Standard deviation
		Binary (yes/no)

Table S1 Aggregation functions used to create additional short-term (day-level) and long-term (admission-level and day-level averages) features of raw ICU time series data.

Squared features
Temperature
Heartrate
pH
Arterial blood pressure (mean, systolic and diastolic)
Bicarbonate
Supplemental oxygen
Leukocytes
Haematocrit
Sodium
Potassium
PaCO2

Table S2 Features that showed a clear non-linear U-shaped relationship with the outcome during visual inspection were squared to created additional features.

Model training

Feature selection was performed using Logistic Regression with an L1 penalty [4, 5]. The regularization parameter was optimized using grid search. Only the features receiving a non-null coefficient were included in the final model. The incidence of our combined endpoint is relatively low compared to an uneventful (good) outcome, also known as an imbalanced dataset. We apply both traditional logistic regression and various machine learning algorithms to study which method leads to the most accurate results for this type of prediction problem. Algorithms were trained to predict the outcome using grid search to optimize hyperparameters (see the Hyperparameter Tuning section and Table S3 and S4).

Hyperparameter Tuning

Model	Hyperparameter	Considered values
All models	Feature selection: C	['0.02', '0.025', '0.03', '0.035', '0.04', '0.045']
	Imputation strategy	['mean', 'median']
Logistic regression	C	['0.001', '0.01', '0.1', '1']
	penalty	['l1', 'l2']
Light GBM	learning_rate	['0.001', '0.005', '0.01', '0.05']
	n_estimators	['100', 500, '1000', '2000']
	subsample	['0.3', '0.4', '0.5', '0.6', '0.7', '0.8', '0.9']
	colsample_bytree	['0.3', '0.4', '0.5', '0.6', '0.7', '0.8', '0.9']
	max_depth	['3', '4', '5', '6', '7']
	min_child_weight	['25', '50', '75', '100']
	scale_pos_weight	['1', '2', '5', '10']
XGBoost	learning_rate	['0.001', '0.005', '0.01', '0.05']
	n_estimators	['100', '250', '500', '1000', '2000']
	reg_alpha	['0', '0.001', '0.01', '0.1']
	reg_lambda	['0', '0.001', '0.01', '0.1']
	subsample	['0.3', '0.4', '0.5', '0.6', '0.7', '0.8', '0.9']
	eval_metric	['error', 'logloss']
	colsample_bytree	['0.3', '0.4', '0.5', '0.6', '0.7', '0.8', '0.9']
	gamma	['0', '0.01', '0.1']
	max_depth	['3', '4', '5', '6', '7']
	min_child_weight	['25', '50', '75', '100']
	scale_pos_weight	['1', '2', '5', '10']
Random Forest	n_estimators	['100', '250', '500', '1000', '2000']
	max_depth	['3', '4', '5', '6', '7']
	min_samples_leaf	['25', '50', '75', '100']
	max_features	['auto', 'log2', 'sqrt']
SVC	kernel	['linear', 'poly', 'rbf']
	degree	['2', '3', '4']
	gamma	['0.001', '0.01', 'scale']
	class_weight	[None, 'balanced']
	C	['0.01', '0.1']

Table S3 Hyperparameters and considered values for grid search optimization for each machine learning classifier

Model	Hyperparameter	Best performing parameter
Logistic regression	Feature selection: C	0.025
	Imputation strategy	median
	C	0.01
	penalty	l2
Light GBM	Feature selection: C	0.025
	Imputation strategy	median
	learning_rate	0.01
	n_estimators	1000
	subsample	0.9
	colsample_bytree	0.3
	max_depth	4
	min_child_weight	25
	scale_pos_weight	1
XGBoost	Feature selection: C	0.045
	Imputation strategy	median
	learning_rate	0.005
	n_estimators	2000
	reg_alpha	0
	reg_lambda	0.1
	subsample	0.8
	eval_metric	error
	colsample_bytree	0.3
	gamma	0.01
	max_depth	6
	min_child_weight	25
	scale_pos_weight	1
Random Forest	Feature selection: C	0.025
	Imputation strategy	median
	n_estimators	100
	max_depth	7
	min_samples_leaf	25
	max_features	sqrt
SVC	Feature selection: C	0.025
	Imputation strategy	median
	kernel	rbf
	degree	2
	gamma	0.001
	class_weight	balanced
	C	0.1

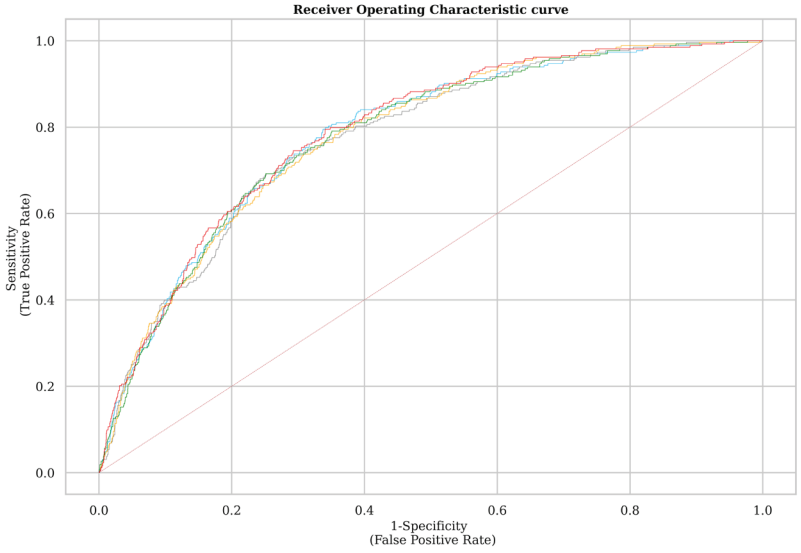
Table S4 Best performing hyperparameters for each machine learning classifier

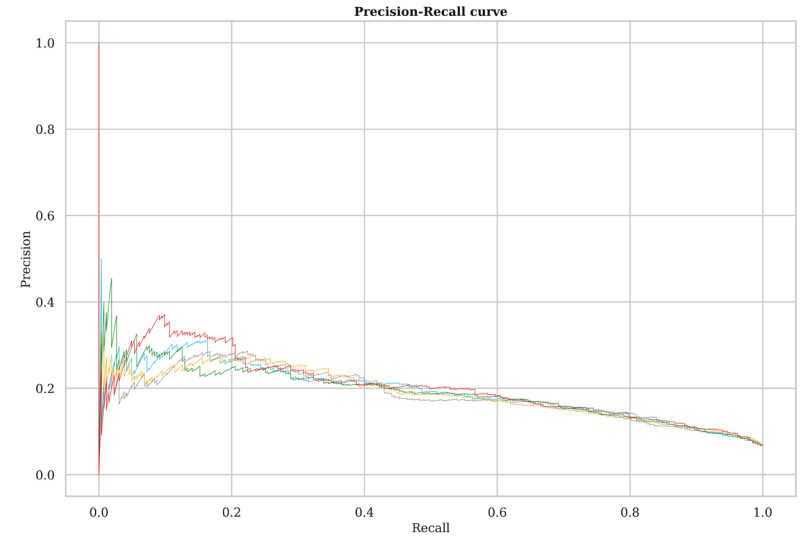
Derivation and validation cohorts

Cohort	Derivation					Validation				
Period	2003-2016					2016-2019				
Characteristic	**Total**	**No events**	**Readmission and/or death**	**Readmission only**	**Death only**	**Total**	**No events**	**Readmission and/or death**	**Readmission only**	**Death only**
ICU admissions, no. (%)	14105 (100)	13354 (94.7)	751 (5.3)	610 (4.3)	173 (1.2)	3929 (100)	3666 (93.3)	263 (6.7)	189 (4.8)	74 (1.9)
Length of stay, days, mean (SD)	3.5 (5.2)	3.3 (5.1)	5.5 (6.4)	5.0 (6.1)	6.8 (7.1)	3.2 (4.5)	3 (4.4)	5 (5.4)	4.8 (5.6)	5.3 (5.1)
Demographics										
Age, years, mean (SD)	63.3 (15.3)	63.2 (15.3)	65.0 (15.5)	63.5 (15.4)	71.5 (13.5)	63.3 (15)	63.1 (15)	66.7 (14.6)	65.5 (15)	69.7 (13.4)
Female, no. (%)	4439 (32.1)	4180 (32.0)	259 (34.9)	206 (34.2)	71 (41.5)	1216 (30.9)	1131 (30.9)	85 (32.3)	57 (30.2)	28 (37.8)

Table S5 Characteristics of derivation and validation cohort.

Model performance





Classifier	AUROC (95% CI)	AUPRC	Number of features
Gradient Boosting	0.780 (0.747-0.814)	0.189	180
Support Vector Classifier	0.789 (0.756-0.822)	0.202	107
Random Forest	0.780 (0.746-0.813)	0.193	104
Logistic Regression	0.783 (0.749-0.816)	0.195	107
Light GBM	0.776 (0.742-0.809)	0.186	169

Figure S1. Receiver operating characteristic curve and precision-recall curve for different algorithms. The AUC of both curves and the number of features used in the models are shown in the lower panel. *AUROC* area under the receiver operating characteristic curve. *AUPRC* area under the precision-recall curve.

Algorithms	Combined outcome			Mortality only model			Readmission only model			Combination of models		
	AUROC	AUPRC	Number of features	AUROC	AUPRC	Number of features	AUROC	AUPRC	Number of features	AUROC	AUPRC	Number of features
XGBoost	0.78 (0.747-0.814)	0.189	180	0.834 (0.777-0.891)	0.086	39	0.715 (0.673-0.757)	0.102	91	0.748	0.161	39+91
Non-linear Support Vector Classifier	0.789 (0.756-0.822)	0.202	107	0.836 (0.779-0.893)	0.108	45	0.734 (0.692-0.775)	0.102	83	0.776	0.191	45+83
Random Forest	0.78 (0.746-0.813)	0.193	104	0.845 (0.789-0.901)	0.092	32	0.737 (0.695-0.778)	0.12	103	0.781	0.192	32+103
Logistic Regression	0.783 (0.749-0.816)	0.195	107	0.849 (0.793-0.904)	0.106	32	0.743 (0.702-0.784)	0.121	83	0.782	0.195	32+83
Light GBM	0.776 (0.742-0.809)	0.186	169	0.839 (0.782-0.895)	0.101	30	0.724 (0.682-0.766)	0.108	77	0.762	0.174	30+77

Table S6 Performance and number of features of the models trained on the combined outcome (mortality and readmission), mortality only, readmission only and a combined prediction of both models using different algorithms and tested on the validation cohort. The prediction of the combination of the separate mortality and readmission models is calculated as: $p = 1 - (1 - p_{\text{mortality}}) \cdot (1 - p_{\text{readmission}})$, where $p_{\text{mortality}}$ is the prediction by the mortality model and $p_{\text{readmission}}$ is the prediction by the readmission model. *AUROC* area under the receiver operated characteristic curve. *AUPRC* area under the precision-recall curve.

Algorithms	Combined outcome			Mortality only model			Readmission only model		
	AUROC	AUPRC	Number of features	AUROC	AUPRC	Number of features	AUROC	AUPRC	Number of features
XGBoost	0.799 (0.783-0.815)	0.179 (0.165-0.193)	180	0.879 (0.861-0.897)	0.158 (0.118-0.198)	39	0.768 (0.761-0.775)	0.122 (0.111-0.133)	91
Non-linear Support Vector Classifier	0.787 (0.773-0.801)	0.176 (0.159-0.193)	107	0.867 (0.851-0.883)	0.143 (0.105-0.181)	45	0.76 (0.743-0.777)	0.107 (0.095-0.119)	83
Random Forest	0.78 (0.765-0.795)	0.165 (0.153-0.177)	104	0.865 (0.85-0.88)	0.145 (0.118-0.172)	32	0.762 (0.748-0.776)	0.113 (0.101-0.125)	103
Logistic Regression	0.786 (0.771-0.801)	0.171 (0.156-0.186)	107	0.872 (0.85-0.894)	0.151 (0.122-0.18)	32	0.755 (0.738-0.772)	0.11 (0.1-0.12)	83
Light GBM	0.795 (0.779-0.811)	0.172 (0.16-0.184)	169	0.872 (0.857-0.887)	0.142 (0.113-0.171)	30	0.765 (0.756-0.774)	0.116 (0.105-0.127)	77

Table S7 Performance of the models trained on the combined outcome (mortality and readmission), mortality only or readmission only using different algorithms **after 10-fold cross-validation**. *AUROC* area under the receiver operated characteristic curve. *AUPRC* area under the precision-recall curve.

Performance metric definitions

		Predicted condition	
		Positive	Negative
Actual condition	Positive	True positives: 189	False negatives: 74
	Negative	False positives: 1091	True negatives: 2575

$$\text{sensitivity (true positive rate, recall)} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} = 0.72$$

$$\text{specificity (true negative rate)} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} = 0.70$$

$$\text{false positive rate} = 1 - \text{specificity} = 0.30$$

$$\text{positive predictive value (precision)} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} = 0.15$$

Figure S2. Confusion matrix, where patients that were readmitted or died within 7 days after ICU discharge were labelled as 'positive' cases. Predictions were performed on the validation cohort (n=3929) using a predicted risk threshold of 6%, with commonly used metrics for evaluating discrimination performance of a classification model. The receiver operating characteristic (ROC) curve plots sensitivity (=true positive rate/recall; y-axis) against (1- specificity) (=false positive rate; x-axis), whereas the precision-recall curve plots positive predictive value (=precision; y-axis) against sensitivity (=recall; x-axis). With imbalanced datasets (i.e. positive << negative) the area under the ROC curve (AUROC) is overly optimistic since specificity is heavily influenced by the large number of negatives in the dataset. In addition, while the baseline AUROC of a random (non-useful) model will be 0.5, the baseline AUPRC equals the incidence rate condition of the validation set, in our case 0.067 (combined outcome readmission and mortality: 6.7%).

The distribution of the predicted outcomes for the XGBoost model is shown in Figure S3. The group with a readmission/mortality event has a much wider predicted probability distribution extending up to almost 0.7 (70%), while in the no event group predicted probabilities are mostly below 0.2 (20%), suggesting that the model is clearly capable of separating the groups. The overlap in the low probability range shows that still a number of patients falsely receive a low probability, even though they have been readmitted and/or died.

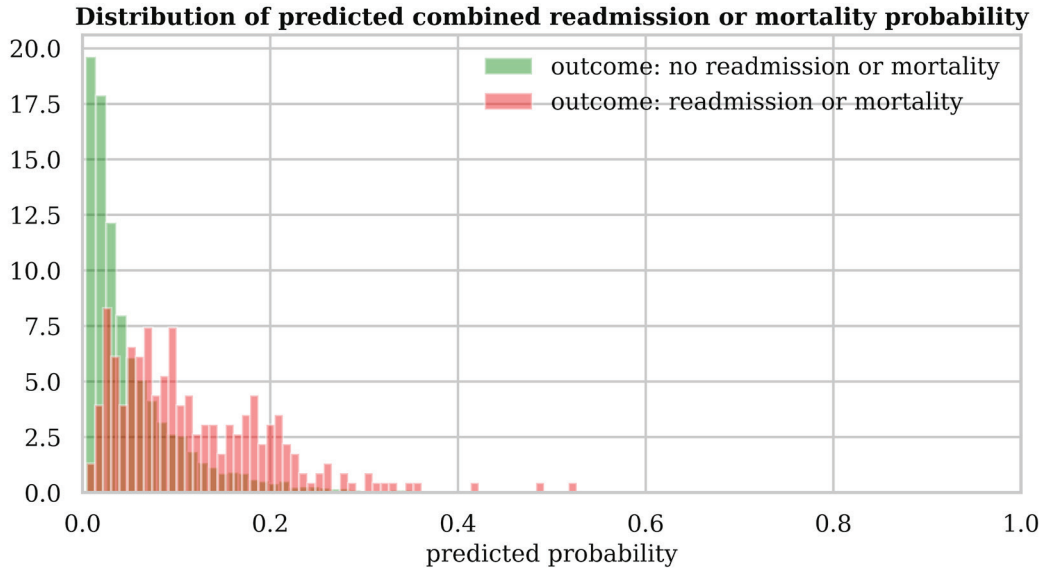


Figure S3 Distribution of the predicted combined readmission and mortality probability for patients in the validation cohort. Patients that have been readmitted or died within seven days are plotted in red while the no event group in green. The two groups are normalized so that the area under each curve is equal to 1, since the number of patients in the readmission/mortality group is much smaller than the no event group. While patients in the readmission/mortality group receive higher probabilities (up to 0.7) than the no event group, there is an overlap between the two groups, suggesting an imperfect classifier.

Decision curve analysis

Decision curve analysis visualizes the so-called net benefit, defined as the difference between the true positives (actual readmissions/deaths) and the false positives (incorrectly identified patients, that could have been discharged), corrected by a factor determined by a threshold the clinician chooses to accept: the readmission probability (see Figure S4).

$$\text{Net benefit} = \frac{\text{True Positives}}{n} - \frac{\text{False Positives}}{n} \cdot \frac{p_t}{1-p_t}$$

Figure S4 Net benefit formula, where *n* is the total number of patients and p_t is the threshold probability. If the predicted probability of our model is higher than the chosen threshold, the patient is classified as positive and we will assume this patient would be readmitted or die if discharged from ICU.

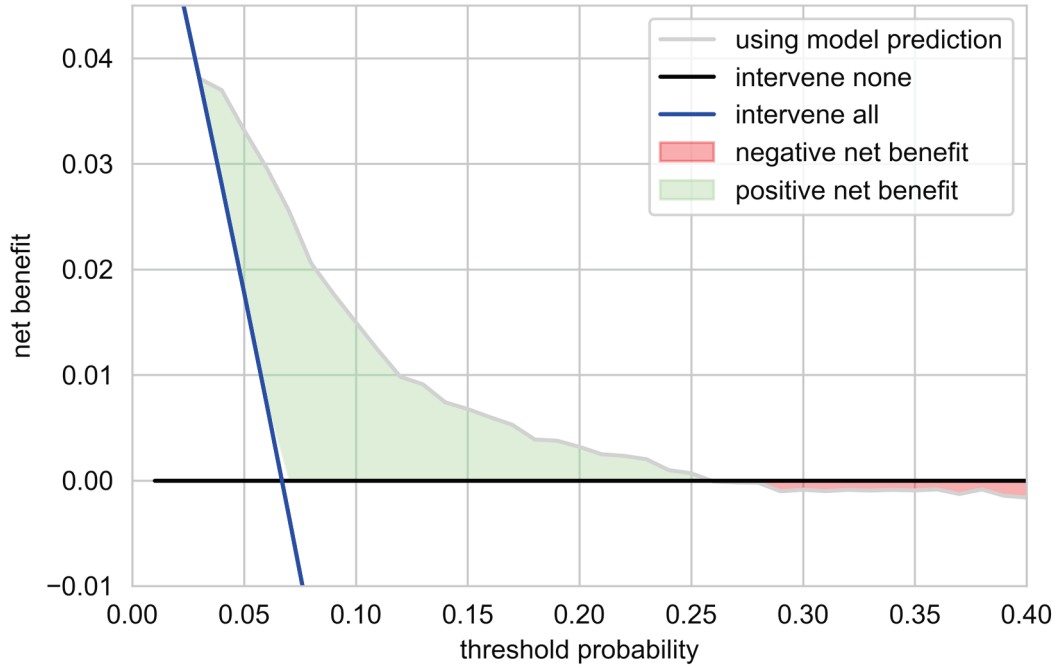


Figure S5 Decision curve analysis for our model showing net benefit for different strategies. The 'intervene none' line shows the strategy of discharging patients based on current practice, which has a net benefit of 0. The 'intervene all' line shows the strategy of postponing discharge from the ICU for all patients, which will only lead to a significant net benefit when using very low thresholds, leading to many unnecessarily prolonged stays. Using our prediction model (grey line), we can demonstrate that for clinically relevant thresholds (~3 to ~30%) net benefit is higher (green area) than the discharge all or none strategies.

Impact analysis

Group	Definition
1. Short-stay	Patients with a length of stay of less than two days
High-risk (> 6%)	Short-stay patient with a combined risk of more than 6% at the moment of discharge
Low-risk (< 6%)	Short-stay patient with a combined risk less than 6% at the moment of discharge
2. Long-stay	Patients with a length of stay of more than two days
Improving	Long-stay patients with a combined risk of more than 6% at the moment of discharge, for whom the risk was at least 2 percentage points higher at the beginning of the previous day
Not improving	Long-stay patients with a combined risk of more than 6% at the moment of discharge, for whom the risk was not at least 2 percentage points higher at the beginning of the previous day
Already low risk	Long-stay patients for whom the combined risk was less than 6% at the beginning of the previous day
Optimal	Long-stay patients for whom the combined risk was less than 6% at the moment of discharge, and more than 6% at the beginning of the previous day
Worsening	Long-stay patients with a combined risk of more than 6% at the moment of discharge, for whom the risk was at least 2 percentage points lower at the beginning of the previous day

Table S8 Definition of the groups for impact assessment. First the patients are divided into (1) short-stay patients and (2) long-stay patients. Short-stay patients are patients with an admission length of less than two days, whereas long-stay patients are patients with an admission length of two or more days.

Group definitions

Since short-stay patients are often low-risk elective surgical patients, we first divided the patients into short and long-stay groups. To further divide the short-stay patients in low and high-risk groups, we chose a threshold of 6%, based on the population risk rate of 5.3% and for the reason that our model has the highest net benefit slightly above this population risk rate (Figure S4). For our analysis, we thus define the following two short-stay groups (Table S8):

High-risk: patients with a combined risk of more than 6% at the moment of discharge and

Low-risk: patients with a combined risk of less than 6% at the moment of discharge.

For long-stay patients, we use readmission probability-time curves to divide them into groups by using the same risk threshold of 6%. Since we expected the *change* in risk to be important for these patients, we arbitrarily chose a change in risk threshold of 2%, thus creating five groups:

Improving: long-stay patients with a combined risk of more than 6% at the moment of discharge, for whom the risk was at least 2 percentage points higher at the beginning of the previous day. It appears these patients are improving in their last admission days.

Not improving: long-stay patients with a combined risk of more than 6% at the moment of discharge, for whom the risk was *not* at least 2 percentage points higher at the beginning of the previous day. Since their combined risk does not decrease substantially in the last days, we expect them not to improve any further when staying longer in the ICU. Potentially, discharging these patients earlier could have been possible, since according to the model these patients had a similar readmission risk at the beginning of the previous day. However, since the readmission rate in this group is relatively high, we chose not to use this in the potential impact calculation.

Already low risk: long-stay patients for whom the combined risk was less than 6% at the beginning of previous day. In the figure with the readmission curves, we see that their risk flattens at the beginning of the previous day, showing that they potentially could have been discharged a day earlier.

Optimal: long-stay patients for whom the combined risk was less than 6% at the moment of discharge, and more than 6% at the beginning of the previous day. It appears these patients have improved in the last days, and have now a relatively low-risk.

Worsening: long-stay patients with a combined risk of more than 6% at the moment of discharge, for whom the risk was at least 2 percentage points lower at the beginning of previous day.

Group	Impact group	Group Size (%)	LOS (days)	Readmission rate (%)	Mortality rate (%)	LOS incl. readmissions (days)	Discharge management	New LOS (days)	Change in LOS (%)
1. Short-stay	High-risk (> 6%)	14.1	1.16	9.4	3.4	1.79	Postpone discharge	2.16	86
	Low-risk (< 6%)	50.5	1.05	2.1	0.3	1.19	Discharge as planned	1.05	0
	Total short-stay	64.6	1.07	3.7	1.0	1.32		1.29	20
2. Long-stay	Improving	7.8	11.22	9.2	5.4	11.85	Postpone discharge	12.22	8.9
	Not improving	8.7	13.81	10.0	4.4	14.49	Discharge as planned	13.81	0.0
	Already low risk	12.2	9.69	4.5	0.1	10.00	Discharge earlier	8.69	-10.3
	Optimal	3.6	9.32	8.5	0.5	9.89	Discharge as planned	9.32	0.0
	Worsening	3.1	11.56	9.7	10.7	12.22	Evaluate and postpone discharge	12.56	8.7
	Total Long-stay	35.4	11.17	7.8	3.3	11.70		11.13	-0.3
Total	*Total*	100.0	4.64	5.1	1.8	4.99	Total	4.77	2.8

Table S9 Descriptive statistics and suggested actions for all distinguished groups. *LOS* Length of Stay. *Change in LOS* refers to the relative increase or decrease in length of stay by changing the discharge management of the specific patient group.

Discharge management

Short-stays

Table S9 shows that the readmission rate is 9.4% for the high-risk short-stays versus 2.1% for the low-risk long-stays. The low-risk group consists of approximately 3 times more patients than the high-risk group. Also the mortality rate is higher in the high-risk group (3.4%) versus the low-risk group 1b (0.3%). The length of stay (LOS) in days is around 1 day for both groups. The length of stay including readmissions is calculated as initial admission length + readmission probability * average duration of a readmission. The average duration of a readmission was 6.8 days in our validation cohort, so much longer than the primary admission. The right side of the table shows the suggested action for the groups. In these calculations, we assume the high-risk group is discharged one day later, which means an increase in admission length of 86% for that group.

Long-stays

The 'improving' group has a relatively high readmission and mortality rate, whereas the 'already low risk' group show a low readmission and mortality rate. The suggested actions are discharging this first group 1 day later; which would mean an increase in length of stay of 8.9%, whereas the latter group could be discharged one day earlier, which would mean a decrease in length of stay of 10.3%. The long-stay 'worsening' group show an increase in the readmission probability and are therefore suggested to be evaluated by an intensivist which may lead to postponing discharge.

Scenarios

For the impact analysis of our model on readmission rates and total length of stay, we performed an analysis using two scenarios. In the first scenario we assumed that the readmission rates of the groups that will be kept in the ICU longer (*short-stay high-risk*, *long-stay improving* and *long-stay worsening*) to drop by 15% (Table S10a). In the second scenario we assume the readmission rates in these groups to drop by 30% (Table S10b). These percentages were based on the drop in the readmission probability of the long-stay *improving* group between the beginning of the previous day and the moment of discharge, which is around 30%. However, this is a critical assumption which requires validation in a prospective trial. We assumed that the suggested action of discharging the *already low risk* group one day earlier does not have a negative impact on readmission rates, since the predicted readmission probability of this group at discharge was similar to the predicted readmission at the

beginning of the previous day. In Table S10 we estimated the effect of this decreasing readmission rate on the *total* length of stay including the readmission. We see that increasing the length of stay of the high-risk short-stays with 1 day (86%) only leads to an increase in *total* length of stay of 45 to 50% if the model has the expected effect on reducing readmission rates. We can conclude from the last rows of Table S10a and Table S10b that the potential impact of the model is 6.9 % to 13.8% reduction in readmission rate with a modest increase in length of stay of 1.6% to 2.1%. Furthermore, when a fraction of the *not improving long-stay patients* which do not seem to improve during the last days of their stay could have been discharged earlier, the use of the model might lead to a lower length of stay as well as a lower readmission rate.

a

Group	Impact group	Readmission rate (%)	Relative risk reduction (%)	LOS including readmissions (days)	Change in LOS incl. readmissions (%)
1. Short-stay	High-risk (> 6%)	8.0	-15	2.70	50
	Low-risk (< 6%)	2.1	0	1.19	0
	Total short-stay	3.4	-8.4	1.52	15
2. Long-stay	Improving	8.0	-15	12.75	8
	Not improving	10.0	0	14.49	0
	Already low risk	4.5	0	9.00	-10
	Optimal	8.5	0	9.89	0
	Worsening	8.3	-15.0	13.12	7
	Total long-stay	7.3	-5.6	11.63	-0.6
Total	*Total*	4.8	**-6.9**	5.09	2.1

b

Group	Impact group	Readmission rate (%)	Relative risk reduction (%)	LOS including readmissions (days)	Change in LOS incl. readmissions (%)
1. Short-stay	High-risk (> 6%)	6.6	-30	2.60	45
	Low-risk (< 6%)	2.1	0	1.19	0
	Total short-stay	3.0	-17	1.49	13
2. Long-stay	Improving	6.5	-30	12.66	7
	Not improving	10.0	0	14.49	0
	Already low risk	4.5	0	9.00	-10
	Optimal	8.5	0	9.89	0
	Worsening	6.8	-30	13.02	7
	Total long-stay	6.9	-11.2	11.60	-0.8
Total	*Total*	4.4	**-13.8**	5.07	1.6

Table S10 Potential impact of the model on readmission rates and *total* length of stay (length of stay including the duration of the readmission). Table S7a assumes a relative reduction of readmission rates in the specified groups of 15% whereas in Table S7b the assumption is a relative reduction of 30%. *Change in LOS incl. readmissions* refers to the relative increase or decrease in total length of stay (e.g. including readmissions) by changing the discharge management of the specific patient group.

Published ICU discharge models

Authors	Year of publication	Patients (n)	Outcome	Incidence (%)	Performance Metric	Method	Number of features
Daly et al.	2001	5,475	Mortality within hospital stay	3.7	ROC AUC = 0.86	Logistic Regression	5
Fernandez et al.	2006	1,156	Mortality within hospital stay	9.6	ROC AUC = 0.88	Logistic Regression	1
Gajic et al.	2008	1,131	Readmission within hospital stay	8.8	ROC AUC = 0.75	Logistic Regression	5
Fernandez et al.	2010	3,587	Mortality within hospital stay	6.7	ROC AUC = 0.84	Logistic Regression	1
Frost et al.	2010	14,952	Readmission within hospital stay	6.6	ROC AUC = 0.66	Logistic Regression	8
Ouanes et al.	2012	3,462	Readmission or mortality within 7 days	3.0	ROC AUC = 0.74	Logistic Regression	6
Badawi and Breslow	2012	469,976	Readmission within 48 hours Mortality within 48 hours	2.5 0.9	ROC AUC = 0.71 ROC AUC = 0.92	Logistic Regression	23 26
Jo et al.	2015	343	Readmission within the same hospital admission	9.6	ROC AUC = 0.76	Logistic Regression	5-6 [a]
Luo et al.	2016	18,412	Mortality within 30 days Mortality within 6 months	3.4 9.5	ROC AUC = 0.860 ROC AUC = 0.842	Support Vector Machine	3-53 [a,b] 35-85 [a,b]
Desautels et al.	2017	2,018	Readmission or mortality within 48 hours	4.4	ROC AUC = 0.71	Gradient Boosting	76
Venugopalan et al.	2017	29,997	Readmission within 30 days	26.0	MCC = 0.73	Combined model LR/ANN/CRF	87
Fabes et al.	2017	4,212	Readmission or mortality within 14 days	8.4	ROC AUC = 0.68	Logistic Regression	8
Rojas et al.	2018	24,885	Readmission within 72 hours	11.4	ROC AUC = 0.76	Gradient Boosting	54-176 [a]
Ng et al.	2018	4,632	Readmission or MET/cardiac arrest call within hospital stay	20.8	ROC AUC = 0.72	Logistic Regression	15
Xue et al.	2019	1,170	Readmission within 30 days	26.5	ROC AUC = 0.66	Logistic Regression	75-185 [c]
McWilliams et al.	2019	9462	Readmission within hospital stay or in-hospital mortality	5.83-12.57 [d]	ROC AUC = 0.888-0.894 [d]	Logistic Regression+ Random Forest	22

Table S11 Published ICU discharge models. *AUC* area under the curve. *LR* logistic regression. *MCC* Matthews correlation coefficient. *ANN* artificial neural network. *CRF* conditional random fields. ROC receiver operated characteristic. [a] Models with varying number of input features presented in manuscript. [b] includes natural language processing. [c] includes subgraph processing. [d] tested on two separate cohorts

User Interface

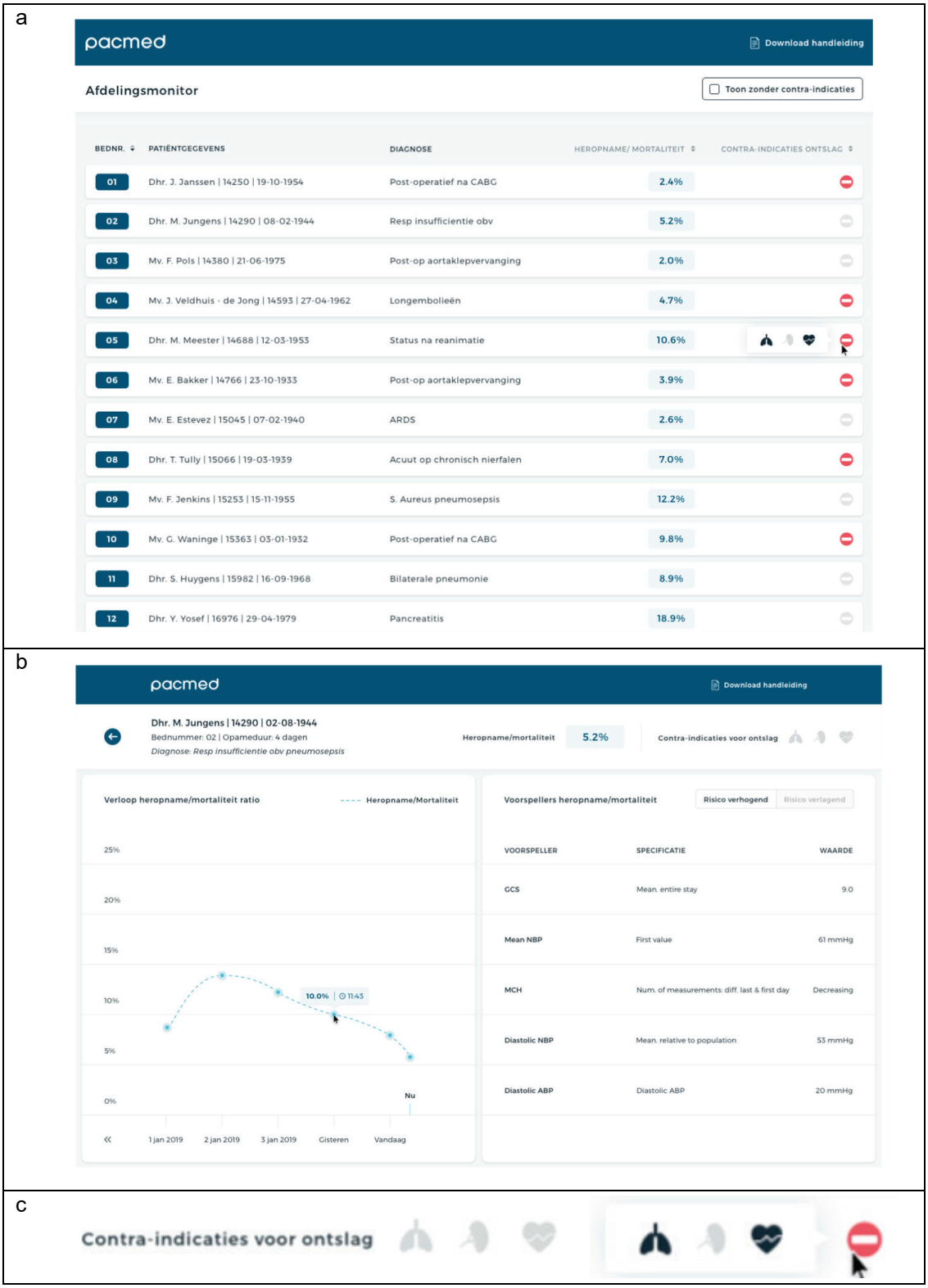


Figure S6 Pre-production version of the Dutch user interface for bedside decision support. Figure S5a: Overview of currently admitted patients, showing location of the patient (bed number), diagnosis, readmission/mortality prediction and contra-indications for discharge (e.g. ventilated). Figure S5b: Detail screen for one of the admitted patients displaying the trend in the predicted risks and the most important predictors. Figure S5c: Enlarged portions of Figure S5b and S5c showing the symbols for conditions that may prevent the discharge of patients (contraindications for discharge, Dutch: *Contra-indicaties voor ontslag)*. Black symbol: patients has contra-indication. Grey symbol: patient does not have that contra-indication. Patient names have been pseudonymized.

References

1. Badawi O, Breslow MJ (2012) Readmissions and Death after ICU Discharge: Development and Validation of Two Predictive Models. PLoS ONE 7:. https://doi.org/10.1371/journal.pone.0048758

2. Fabes J, Seligman W, Barrett C, et al (2017) Does the implementation of a novel intensive care discharge risk score and nurse-led inpatient review tool improve outcome? A prospective cohort study in two intensive care units in the UK. BMJ open 7:e018322. https://doi.org/10.1136/bmjopen-2017-018322

3. Rojas JC, Carey KA, Edelson DP, et al (2018) Predicting intensive care unit readmission with machine learning using electronic health record data. Annals of the American Thoracic Society 15:846–853. https://doi.org/10.1513/AnnalsATS.201710-787OC

4. Tibshirani R (1996) Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58:267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

5. Talenti L, Luck M, Argy N, et al (2015) L 1 Penalized logistic Regression Feature Preselection for Classification Tree: Application to the Diagnostic of Severe Imported Malaria Based on Heterogeneous Data. JSM Math Stat 2:1012