Length-of-stay and mortality prediction for a major hospital through interpretable machine learning

Dimitris Bertsimas\textsuperscript{1}, Jean Pauphilet\textsuperscript{1}, Jennifer Stevens\textsuperscript{2}, Manu Tandon\textsuperscript{2}

\textsuperscript{1}Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA
\textsuperscript{2}Beth Israel Deaconess Medical Center, Boston, MA.

Abstract

Importance: Understanding the discharge process at a hospital level is key in improving efficiency and quality of care.

Objective: Investigate how machine learning can help anticipate various aspects of patient discharges, from predicting length-of-stay to discharge destination or hospital mortality.

Design: Retrospective study performed on inpatients admitted at Beth Israel Deaconess Medical Center between January 2017 and August 2018.

Setting: Single-center study in a large academic medical center in the Boston area.

Participants: We included inpatients admitted at BIDMC between January 2017 and August 2018, excluding patients admitted into psychiatry, obstetrics and newborns. The final cohort consisted of 48,770 unique admissions (32,829 unique patients).

Main outcomes and measures: We predicted whether a patient will be discharged in the next 24 or 48 hours, whether she will stay more than 7 or 14 days and predict discharge destination among home, home with services, extended care facility and hospital mortality. Data is collected retrospectively from electronic health records.

Methods and results: We used data from 48,770 admissions at BIDMC (48.6\% female, median age 65 years old, median length-of-stay 4.15 days) to answer four length-of-stay-related questions, as well as to predict discharge destination. We applied five different machine learning algorithms. With the best performing method, we predict same-day discharges (remaining length-of-stay < 1 day) with an area under the receiving operator curve (AUC) of 0.962 (95\%CI 0.960-0.964), next-day discharges (remaining length-of-stay < 2 days) with an AUC of 0.858-0.865, long-stay patients (overall length-of-stay > 7/14 days) with an AUC of 0.804-0.814 and 0.797-0.814 respectively. Similarly, we accurately predict discharge destination (weighted AUC of 0.829-0.835), hospital mortality (AUC 0.957-0.965) and discharge to extended care facility (AUC 0.860-0.866).

Conclusions: We are able to accurately identify same-day or next-day discharges, long-stay patients and predict discharge destination. Though less accurate, simpler and interpretable models, such as decision trees, demonstrate very good predictive power, provide insights on discharge barriers and have been instrumental in interacting with care providers. In addition, those models are, compared to deep learning approaches, frugal in data and computational power and provide production-level analytics for EHRs.

1 Introduction

Patient discharges drive bed availability, which is one of the most critical hospital resources. From a clinical perspective, prolonged length-of-stay is associated with negative outcomes for the patient, such as increased infection risks. From an operational perspective, length-of-stay is a meaningful indicator of costs, resources consumption and patient flows. As a result, accurate length-of-stay
prediction of inpatients could improve care delivery at a patient level, by highlighting current discharge barriers or anticipating long stays, and at a healthcare facility level, through improved resource management and planning. Discharge destination is another important factor of the discharge process. Indeed, further case management resources should be allocated to patients requesting a bed in extended care facilities, while high-mortality-risk patients should be detected early on by the clinicians. In this work, we combine advanced machine learning methods with expertise and insights from medical staff, social workers and operations to provide quantitative estimates of length-of-stay and discharge destination, with hospital mortality as a special case.

**Related work and contributions:** Being a surrogate for negative clinical outcomes as well as operational performance, length-of-stay has received a vivid interest in the academic literature, often in combination with hospital mortality (Awad et al. 2017). High-quality data available from electronic health records (EHR) as well as recent successes in machine learning have fostered new approaches on the matter, to which our work belongs: with a focus on prediction rather than causality, we use machine learning techniques to predict length-of-stay and discharge destination for a general inpatient population. Recently, Rajkomar et al. (2018b) proposed a data processing pipeline to convert raw EHR data into a standardized output and use state-of-the-art deep learning models to predict hospital mortality, readmissions and long stays. Despite excellent predictive power and extensive human and computational resources, their study is only retrospective and has not been integrated within any EHR system. Indeed, they rely on medical notes which might not be available in real-time and raise data privacy issues, especially if third-party computational resources are needed. The black-box nature of their models also impedes adoption from doctors and caregivers. Finally, in our view, the gap between their deep learning approach and simple linear baselines (Rajkomar et al. 2018a) is not significant enough to justify such a heavy implementation requirement.

**Contributions and structure:** In this work, we propose a simple expertise-driven modeling framework to capture the state of each inpatient as she stays in the hospital, instead of a deep learning approach. Compared to previous work (Rajkomar et al. 2018b), we use a hospital-centric rather than patient-centric time scale and only leverage features which are reliably available after admission, on a daily basis. Second, from this unique framework, we apply a broad collection of machine learning techniques to answer four length-of-stay-related questions: identify same-day and next-day discharges and predict more-than-7 and more-than-14-day stays. We then investigate the question of predicting discharge destination among home, home with services, extended care facility and death. For all tasks, we match or surpass state-of-the-art methods, even without using raw medical notes. Ensemble methods are the most accurate, but linear models and decision trees provide very good predictive power, together with actionable insights to practitioners thanks to their interpretability. Finally, our work illustrates that emphasis on modeling and interpretability does not induce poor predictive accuracy, but achieves higher engagement from the clinicians and care providers, lower data requirement and light computational effort. As a result, we were able to conduct the project from initial data exploration to production-level deployment in less than twelve months.

## 2 Methods

We applied the following methodology.
2.1 Study population

We gathered data from the EHR of inpatients admitted at BIDMC between January 2017 and August 2018. We excluded patients admitted into psychiatry, obstetrics and newborns. The final cohort consisted of 48,770 unique admissions (32,829 unique patients), whose demographics and relevant variables are summarized in Table 1. The dataset contained patient demographics, provider orders, ICD10 diagnosis codes from previous and current admissions, medications, blood laboratory values, vital signs and key scores (e.g., pain scale, mobility score). Institutional review board at BIDMC approved the study with waiver of informed consent.

2.2 Modelling

Previous approaches (Rajkomar et al. 2018b) considered a patient-centric time scale. For each patient, the clock starts at admission and a prediction can be made in the following 12 or 24 hours. On the contrary, we adopted a hospital-centric time scale, where predictions are made on a fixed schedule (daily at 11:59pm in our study) for all inpatients. In our view, this fixed schedule mimics the reality of inpatient management and better captures the periodicity of hospital operations.

We defined an observation as the state of a patient at the end of each day. Consequently, the final dataset contained 347,692 unique observations. Most variables were counting operations (e.g., number of blood bank orders submitted in the past day or since admission). For lab results and vitals, we computed daily average, amplitude and trend.

2.3 Missing data

Most of the data consisted of records of events (e.g., record for surgeries, medical orders, medications) and therefore could not be missing. Some variables, such as homelessness indicator, were mostly missing and equal to Yes otherwise, so we considered missing as its own category (No in this case). For each patient, lab results and vital signs were imputed using linear interpolation when possible, given their times series nature. Finally, we imputed the remaining missing values using an optimization-based imputation method (Bertsimas et al. 2018).

For production deployment, we favored tree-based methods for they can handle missing data by applying mode imputation: At each split, a rule is applied to decide whether to go on the right or left side of the tree. If information required to compute the rule is missing, one decides to go in the direction where the majority of the training observations went. For instance, at Leaf 1 in Figure 1, a prediction is made based on the value of the daily platelet level of the patient. Conditionally on being in that part of the tree, we historically observed 89 patients with platelet level lower than 200.5, and 109 patients with higher platelet levels. Without any information on the platelet level for a particular patient, we would then assume her platelet level is higher than 200.5, as for the majority of past patients.

Note that, apart from regression, all methods in our study are tree-based methods. However, while CART (Breiman et al. 1984) and Optimal Trees produce (Bertsimas and Dunn 2017) a single decision tree, gradient boosted trees (Friedman 2001) and random forest (Breiman 2001) combine predictions from multiple trees, hence losing in interpretability.

2.4 Relevant outcomes

We answered four length-of-stay-related questions. For resource planning purposes, predicting same-day or next-day discharges provides visibility on future bed availability. Accordingly, we built models to predict whether the remaining length-of-stay was less than 1 or 2 days. From a
clinical perspective, it is also useful to identify long-stay patients, i.e., patients whose overall length-of-stay exceeds 7 or 14 days. Since early detection of long-stay patients can be crucial, we built classification models using only the first four days of each admission to answer those two questions. We then investigated discharge destination prediction, formulated as a four-class classification problem (between Home, Home with services, Extended Care Facility, Death), with a focus on hospital mortality and extended care facility specifically.

2.5 Model evaluation and statistical analysis

Patients were split based on their admission date into train (Jan - Dec 2017, 60%), validation (Jan - Apr 2018, 20%) and test (May - Aug 2018, 20%) sets. Models were trained on the train and validation sets, using the validation set to calibrate hyper parameters and avoid overfitting. Observations in the train set were weighted according to their class prevalence to account for unbalanced outcomes. We report accuracy metrics computed on the test set only. We assessed performance of classification models by calculating the area under the receiver operating curve (AUC). 1,000 bootstrapped samples were used to calculate 95% confidence intervals.

2.6 Computing resources

Data preprocessing was done on our partner hospital database, natively in SQL. Train and test of the predictive models were done in Python 3.5.2 and Julia 1.0.1 on a MacBook Pro with 2.5 GHz Intel Core i7 CPU and 16GB of RAM.

3 Results

Results are reported in Tables 2 and 3.

3.1 Predicting imminent discharges

For operational purposes, identifying imminent discharges helps predict future hospital census and bed availability.

With all methods, we detected same-day discharges with very high accuracy (AUC above 92%), with gradient boosted trees being the most accurate method in our study (95% CI: 0.960-0.964). As a result, we could predict daily number of discharges with a median absolute error (MAE) of 5 beds only (IQR: 2 beds - 10 beds). Using a random forest, we predicted next-day discharges with an AUC of 0.862 (0.858-0.865).

Comparison with similar studies is always difficult because of different patient population and accuracy metrics. Yet they provide interesting reference points. For surgical patients only, Zanger (2018) developed a deep-learning approach which predicted 24-hour discharges with an AUC of 0.840 (± 0.07), which is outperformed by all methods in our study. Though on a different inpatient population, their work surpassed in accuracy the random forest model from Barnes et al. (2016). Recently, Van Walraven and Forster (2017) use non-parametric models to predict the daily number of hospital discharges with accuracy roughly comparable to ours (median relative error: 1.4%; IQR -5.5% to 7.1%). McCoy et al. (2018) forecast discharge volume in two hospitals using time-series algorithm with a mean absolute error of 11.5 and 11.7 days respectively.
3.2 Anticipating long stays

Long stay patients are typically patients with more complex medical or social conditions and consume a large amount of hospital resources so that identifying them early in their stay could be extremely beneficial. As in Rajkomar et al. (2018b) we identified patients with an overall length-of-stay above 7 days. They reported accuracy 24 hours after admission, which approximately corresponds to prediction after two days at the hospital with our modeling choice. In our two works, the logistic regression model already performed well: we reached an AUC of 0.738 (0.727-0.75) after one day and 0.792 (0.781-0.802) after two days, which was comparable with their adaptation of the logistic model from Liu et al. (2010) (95% CI 0.73-0.77). Having implemented a similar linear benchmark, we were thus optimistic about the comparability of our results. Gradient boosted trees achieved 0.804 (0.793-0.813) after two days, 0.810 (0.804-0.814) overall, which was comparable with their deep learning approach, without medical notes.

Similarly, we trained models to detect overall length-of-stay above 14 days, a threshold more relevant to our partner hospital, and reached a 0.806 AUC (0.797-0.814) using gradient boosted trees. Fortunately, as shown on Table 2, accuracy level globally increases as patients stay longer and more data is collected. After only two days, all methods reached 95% of their full potential.

3.3 Predicting discharge destination

Discharge destination can sometimes be as important as time-to-discharge itself. As presented in Table 3, we were able to accurately classify patients between the four potential discharge destinations with a weighted AUC around 75% for decision trees and 82% for ensemble methods.

Two destinations require the utmost attention: Extended care facility, because they trigger additional case management effort, and death. We built one-versus-all classifiers for those destinations specifically and reached an AUC of 0.863 (0.86-0.866) and 0.961 (0.957-0.965) respectively.

For extended care facility, the AUC after one day only is already very close to the overall AUC (within 1-4% depending on the method), which would enable anticipating administrative bottlenecks early in the stay.

For hospital mortality, our approach outperformed previous studies (Tabak et al. 2014) and even competed with the deep learning model from Rajkomar et al. (2018b) (95% CI: 0.920.96). We noticed that the most predictive factor was the presence of a Do Not Resuscitate document in the patients EHR. One might argue that this feature is too related with the output of interest to be included in the model. To this end, we also trained models which did not include this variable and still demonstrated around 90% AUC.

4 Discussion

In this study, we demonstrated how tailored modeling can be used in combination with interpretable machine learning techniques to provide accurate predictions of significant length-of-stay-related indicators and discharge destinations.

In a hospital environment under increasing financial and operational stress, improvement in care delivery would require “the utilization of advanced data analytics to [...] forecast patient demand patterns, and match capacity and demand” (Rutherford et al. 2017). In this regard, predicted length-of-stay could act as a unified prioritization tool across the hospital, as well as anticipate future capacity and identify complex cases early in their stay. For a large hospital like BIDMC, reducing average length-of-stay, even by 1%, would approximately increase the number of patients seen in a year by 500.
Besides predictive power, our work demonstrates the value of careful modeling and interpretability. Rajkomar et al. (2018b) proposed a generic data processing pipeline to alleviate the burden of data cleaning and modeling, which they consider as the major implementation bottleneck. Our experience has been different. Their pipeline, though viable for a retrospective study, requires immoderate computing power and daily available medical notes. We show that, even without text notes, a tailored modeling can incorporate clinical expertise from physicians, nurses and case managers to reach comparable performance and increased engagement. Our models requires standard computing resources and could easily be recalibrated on a regular basis to adapt to interventions, such as new discharge protocols. Furthermore, as opposed to tree-based methods, deep learning approaches are not interpretable by design. Rajkomar et al. (2018b) use attribution methods to highlight the elements from a patient’s EHR which impacted prediction. This visualization procedure required retraining a model on a restricted version of the data, so the depicted model differs from the one making the actual predictions.

Our study has several limitations. For now, our analysis is single-center only and we are working with a second hospital to adapt and validate the benefits of our approach. Second, any statistical model is only as good as the data it is trained on. Consequently, when predicting time-to-discharge, our models incorporated operational inefficiencies. In presence of undesirable readmissions or administrative delays, it is unclear whether length-of-stay alone is a relevant outcome to consider. Yet, from an operation standpoint, length-of-stay remains a key metric and interpretability of the models allow clinicians to read predictions with a pinch of salt. In addition, our approach brings a valuable piece to the overall length-of-stay vs. readmission risk trade-off. Thirdly, operational deployment of machine learning methods on EHR data requires advanced electronical data collection and management, and reduces the scope of usable information, such as medical notes.

As far as predictive accuracy is concerned, ensemble methods, namely random forest and gradient boosted trees, are the best performing methods. Predicting long stays seemed notably harder than predicting imminent discharges, which could be explained by the fact that patients need to reach fairly standardized milestones in order to be discharged while complex cases are a very diverse population. Yet, simple linear regression is already within 4% from the best performing method. In our opinion, the value of machine learning is to be found elsewhere than plain accuracy.

Linear regression and ensemble methods, like neural networks, can only provide a list of variables ranked by their relative importance. Table 4 presents the five most important variables for predicting whether the overall length-of-stay will be above 7 days using a random forest. In particular, this information is not patient-specific and does not capture non-linear effects. Decision trees, on the other hand, explicit the underlying mechanisms guiding the prediction, while demonstrating satisfactory accuracy. For each patient, a tree provides not only a prediction, but also the path leading to that prediction, which informs both statistical modeling and the care provider: Figure 2 displays the optimal tree predicting hospital mortality with a 0.942 AUC. From this tree, we were able to pinpoint the major predictive power of the Do Not Resuscitate variable and question the use of this variable in our final model. On a similar note, Figure 1 displays particular leaves of the tree predicting long-stay patients, highlighting some clinical (low platelet level, high systolic blood pressure) as well as operational (low discharge volume on Friday-Sunday) barriers to discharge. Of course, those leaves are only valid considering the whole path which leads to them, so conclusions based on those observations should not be drawn too quickly. Interpretability is also a useful tool to avoid bias (Gianfrancesco et al. 2018).
5 Conclusion

To the best of our knowledge, our study is first of its kind to (a) address the length-of-stay and discharge destination prediction task for such a generic inpatient population with a unified data modeling and processing, (b) achieve state-of-the-art accuracy with a broad collection of models, including interpretable ones, (c) be fully integrated into the EHR system of a major hospital, thus demonstrating how powerful analytics can concretely impact care delivery. Yet, further work is needed to ensure and validate that accurate predictions translate into improved quality of care.

Table 1: Summary statistics for the study population in the training and validation (Jan 2017- Apr 2018) and test (May 2018 - Aug 2018) data set.

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Training and validation $(n = 38,999)$</th>
<th>Test $(n = 9,771)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, median (IQR) y</td>
<td>65 (23)</td>
<td>66 (22)</td>
</tr>
<tr>
<td>Female sex, no. (%)</td>
<td>19,357 (49.6%)</td>
<td>4,899 (50.1%)</td>
</tr>
<tr>
<td>Homeless, no. (%)</td>
<td>31 (0.1%)</td>
<td>6 (0.1%)</td>
</tr>
<tr>
<td>Not english speaking, no. (%)</td>
<td>3,921 (10.1%)</td>
<td>1,010 (10.3%)</td>
</tr>
<tr>
<td>Body mass index, median (IQR)</td>
<td>27.35 (8.63)</td>
<td>27.35 (8.49)</td>
</tr>
<tr>
<td>Pre-existing comorbidities, median (IQR)</td>
<td>1 (5)</td>
<td>2 (6)</td>
</tr>
</tbody>
</table>

Previous hospitalization in the last 6 months, no. (%)

| 0 hospitalization             | 26,640 (68.3%)                         | 6,642 (68%)       |
| 1 hospitalization             | 6,779 (17.4%)                          | 1,734 (17.7%)     |
| 2+ hospitalizations           | 5,580 (14.3%)                          | 1,395 (14.3%)     |

Discharge destination, no. (%)

| Expired                      | 1,153 (3.0%)                           | 283 (2.9%)        |
| Extended care facility       | 8,740 (22.7%)                          | 2,114 (21.9%)     |
| Home                        | 16,907 (43.9%)                         | 4,191 (43.5%)     |
| Home with services           | 11,731 (30.4%)                         | 3,055 (31.7%)     |

Primary outcomes

| LOS, median (IQR) d          | 4.15 (4.50)                            | 4.17 (4.46)       |
| overall LOS $\geq$ 7 days, no. (%) | 9,971 (25.6%)                        | 2,485 (25.4%)     |
| overall LOS $\geq$ 14 days, no. (%) | 3,226 (8.3%)                          | 709 (7.3%)        |
Table 2: Summary of the results on predicting length-of-stay (overall and remaining) for logistic regression (LR), CART decision trees (CART), optimal trees with parallel splits (OT), random forest (RF) and gradient boosted trees (GBT). MAE = Median Absolute Error. MRE = Median Relative Error

<table>
<thead>
<tr>
<th>Classification: remaining length-of-stay &lt; 1 day</th>
<th>LR</th>
<th>CART</th>
<th>OT</th>
<th>RF</th>
<th>GBT</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC (95% CI)</td>
<td>0.937 (0.934-0.939)</td>
<td>0.931 (0.929-0.934)</td>
<td>0.935 (0.933-0.938)</td>
<td>0.954 (0.951-0.956)</td>
<td><strong>0.962 (0.96-0.964)</strong></td>
</tr>
<tr>
<td>MAE in # daily discharges (IQR), no.</td>
<td><strong>5.0 (2.0 - 9.0)</strong></td>
<td><strong>5.0 (2.0 - 9.0)</strong></td>
<td>6.0 (3.0 - 14.0)</td>
<td><strong>6.0 (3.0 - 12.0)</strong></td>
<td><strong>5.0 (2.0 - 10.0)</strong></td>
</tr>
<tr>
<td>MRE in # daily discharges (IQR), %</td>
<td><strong>7.7 (3.6 - 15.9)</strong></td>
<td><strong>7.7 (3.6 - 15.9)</strong></td>
<td>9.9 (4.7 - 22.1)</td>
<td>10.7 (4.7 - 19.0)</td>
<td>8.4 (3.9 - 16.8)</td>
</tr>
</tbody>
</table>

| Classification: remaining length-of-stay < 2 days |
|-----------------------------------------------|----|------|----|----|-----|
| AUC (95% CI)                                  | 0.847 (0.844-0.851) | 0.83 (0.827-0.834) | 0.828 (0.824-0.832) | **0.862 (0.858-0.865)** | 0.861 (0.857-0.864) |

| Classification: overall length-of-stay < 7 days |
|-----------------------------------------------|----|------|----|----|-----|
| AUC (95% CI)                                  | 0.798 (0.793-0.803) | 0.752 (0.747-0.758) | 0.768 (0.762-0.773) | 0.809 (0.804-0.814) | **0.81 (0.804-0.814)** |
| AUC at day 1 (95% CI)                          | 0.738 (0.727-0.75) | 0.697 (0.686-0.708) | 0.721 (0.71-0.733) | **0.767 (0.756-0.778)** | 0.755 (0.744-0.766) |
| AUC at day 2 (95% CI)                          | 0.792 (0.781-0.802) | 0.738 (0.727-0.749) | 0.754 (0.743-0.765) | 0.801 (0.791-0.811) | **0.804 (0.793-0.813)** |

| Classification: overall length-of-stay < 14 days |
|-----------------------------------------------|----|------|----|----|-----|
| AUC (95% CI)                                  | 0.797 (0.788-0.806) | 0.728 (0.718-0.737) | 0.746 (0.736-0.755) | 0.795 (0.786-0.803) | **0.806 (0.797-0.814)** |
| AUC at day 1 (95% CI)                          | 0.768 (0.751-0.787) | 0.693 (0.672-0.713) | 0.737 (0.718-0.757) | 0.759 (0.74-0.778) | **0.783 (0.766-0.801)** |
| AUC at day 2 (95% CI)                          | 0.805 (0.788-0.821) | 0.726 (0.707-0.745) | 0.745 (0.725-0.767) | 0.797 (0.779-0.813) | **0.806 (0.791-0.822)** |
Table 3: Summary of the results on predicting discharge destination for logistic regression (LR), CART decision trees (CART), optimal trees with parallel splits (OT), random forest (RF) and gradient boosted trees (GBT).

<table>
<thead>
<tr>
<th>Discharge Destination</th>
<th>LR</th>
<th>CART</th>
<th>OT</th>
<th>RF</th>
<th>GBT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted AUC (95% CI)</td>
<td>0.574 (0.57-0.578)</td>
<td>0.76 (0.756-0.763)</td>
<td>0.749 (0.746-0.753)</td>
<td><strong>0.832 (0.829-0.835)</strong></td>
<td>0.829 (0.827-0.832)</td>
</tr>
<tr>
<td>Weighted AUC at day 1 (95% CI)</td>
<td>0.603 (0.595-0.612)</td>
<td>0.765 (0.757-0.774)</td>
<td>0.747 (0.739-0.756)</td>
<td>0.828 (0.821-0.835)</td>
<td><strong>0.829 (0.822-0.836)</strong></td>
</tr>
<tr>
<td>Weighted AUC at day 2 (95% CI)</td>
<td>0.615 (0.606-0.623)</td>
<td>0.778 (0.77-0.786)</td>
<td>0.75 (0.742-0.759)</td>
<td><strong>0.841 (0.834-0.847)</strong></td>
<td>0.841 (0.834-0.847)</td>
</tr>
</tbody>
</table>

Mortality (with DNR indicator)

<table>
<thead>
<tr>
<th></th>
<th>LR</th>
<th>CART</th>
<th>OT</th>
<th>RF</th>
<th>GBT</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC (95% CI)</td>
<td>0.95 (0.945-0.956)</td>
<td>0.923 (0.915-0.932)</td>
<td>0.942 (0.935-0.948)</td>
<td><strong>0.961 (0.957-0.965)</strong></td>
<td>0.915 (0.904-0.924)</td>
</tr>
<tr>
<td>AUC at day 1 (95% CI)</td>
<td>0.943 (0.916-0.964)</td>
<td>0.895 (0.855-0.93)</td>
<td>0.901 (0.862-0.934)</td>
<td><strong>0.949 (0.924-0.967)</strong></td>
<td>0.93 (0.898-0.957)</td>
</tr>
<tr>
<td>AUC at day 2 (95% CI)</td>
<td>0.938 (0.906-0.961)</td>
<td>0.903 (0.864-0.937)</td>
<td>0.912 (0.875-0.943)</td>
<td><strong>0.952 (0.931-0.969)</strong></td>
<td>0.914 (0.871-0.948)</td>
</tr>
</tbody>
</table>

Mortality (without DNR indicator)

<table>
<thead>
<tr>
<th></th>
<th>LR</th>
<th>CART</th>
<th>OT</th>
<th>RF</th>
<th>GBT</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC (95% CI)</td>
<td>0.871 (0.86-0.88)</td>
<td>0.848 (0.838-0.858)</td>
<td>0.856 (0.846-0.865)</td>
<td><strong>0.91 (0.903-0.917)</strong></td>
<td>0.89 (0.882-0.899)</td>
</tr>
<tr>
<td>AUC at day 1 (95% CI)</td>
<td><strong>0.883 (0.853-0.909)</strong></td>
<td>0.794 (0.751-0.833)</td>
<td>0.814 (0.775-0.846)</td>
<td>0.873 (0.84-0.901)</td>
<td>0.867 (0.835-0.896)</td>
</tr>
<tr>
<td>AUC at day 2 (95% CI)</td>
<td>0.886 (0.847-0.918)</td>
<td>0.801 (0.757-0.84)</td>
<td>0.811 (0.768-0.848)</td>
<td><strong>0.893 (0.862-0.92)</strong></td>
<td>0.871 (0.834-0.903)</td>
</tr>
</tbody>
</table>

Extended Care Facility

<table>
<thead>
<tr>
<th></th>
<th>LR</th>
<th>CART</th>
<th>OT</th>
<th>RF</th>
<th>GBT</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC (95% CI)</td>
<td>0.861 (0.858-0.864)</td>
<td>0.817 (0.813-0.821)</td>
<td>0.821 (0.817-0.825)</td>
<td><strong>0.863 (0.86-0.866)</strong></td>
<td>0.861 (0.858-0.864)</td>
</tr>
<tr>
<td>AUC at day 1 (95% CI)</td>
<td>0.867 (0.857-0.876)</td>
<td>0.829 (0.818-0.839)</td>
<td>0.841 (0.83-0.852)</td>
<td>0.869 (0.86-0.877)</td>
<td><strong>0.875 (0.867-0.885)</strong></td>
</tr>
<tr>
<td>AUC at day 2 (95% CI)</td>
<td>0.882 (0.872-0.891)</td>
<td>0.842 (0.832-0.852)</td>
<td>0.849 (0.839-0.859)</td>
<td>0.881 (0.872-0.889)</td>
<td><strong>0.886 (0.877-0.894)</strong></td>
</tr>
</tbody>
</table>
Figure 1: Look-up at a decision tree for predicting whether a patient will stay more than 7 days on overall. Class 0 corresponds to long stays, while class 1 corresponds to less-than-7-day stays. Leaf 1 suggests that patients with low platelet level (platelet_level < 200.5 k/µL) experience longer stays. Leaf 2 identifies patients with higher-than-normal systolic blood pressure (sys_bp_last > 123.4mmHg) as more likely to experience long stays. Leaf 3 confirms that end of the week (when day of the week is Friday, Saturday or Sunday) is less prone to discharges.
Table 4: Top 5 variables selected by the random forest classifier for predicting whether a patient will stay more than 7 days on overall.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>orders_count_cum</td>
<td>Number of orders placed since admission</td>
</tr>
<tr>
<td>2</td>
<td>lab_orders</td>
<td>Number of lab orders placed in the last 24 hours</td>
</tr>
<tr>
<td>3</td>
<td>mars_count</td>
<td>Number of MARS documentation written in the last 24 hours</td>
</tr>
<tr>
<td>4</td>
<td>abn_b</td>
<td>Number of abnormal blood results received in the last 24 hours</td>
</tr>
<tr>
<td>5</td>
<td>orders_types_cum</td>
<td>Number of unique order types placed since admission</td>
</tr>
</tbody>
</table>

Figure 2: Decision tree predicting inpatient mortality. The method identifies 6 relevant variables.  

- **dnr_ind**: Indicates whether the patient signed a Do Not Resuscitate form (Y/N)  
- **orders_count_cum**: Number of medical orders placed for the patient since admission  
- **abn_b**: Number of abnormal blood tests received within the past 24 hours  
- **hosp_svc**: Hospital service the patient is in  
- **days_inicu**: Number of days spent in the ICU since admission  
- **bilirubin_measured**: Indicates whether bilirubin level is measured (0: not measured, 1: measured and normal level, 2: measured and abnormal level)
References


Carl Van Walraven and Alan Forster. The TEND (Tomorrow’s Expected Number of Discharges) Model Accurately Predicted the Number of Patients Who Were Discharged from the Hospital the Next Day. *Journal of Hospital Medicine*, 13(3):1–6, 2017. doi: 10.12788/jhm.2802.

A Online Supplement: feature definition

In this supplement, we describe in detail the features we created from the raw EHR data to compute our predictions.

We divided the variables into two categories:
1. static variables, which are not supposed to change as the patient stays at the hospital,
2. dynamic variables, which are daily updated.

A.1 Static variables

Static variables represent a patient's socio-economic situation, as well as existing conditions. We assume these variables do not evolve over a patient's stay. We acknowledge that this assumption might not always hold in practice. For instance, for certain conditions, weight can drastically change over one's stay. In addition, some data such as insurance information might not be available and entered in the EHR system directly at admission.

Static variables are obtained from three sources, namely
- Admission data,
- ICD10 codes from previous admissions (billing data),
- Initial Patient Assessment information (completed on the day of admission),

and consist of
- Admission data
  - Patients type and source
  - Hospital service responsible for the patient
  - Age
  - Gender
  - Weight, height, BMI
  - Insurance indicators
  - Homelessness indicator
  - Income category based on ZIP code
  - English-speaking indicator
  - Information about previous admission in the past year (frequency, previous length-of-stay)
- ICD10 codes from previous admissions (billing data)
  - Indicator for each comorbidity, based on Elixhauser comorbidity software [https://www.hcup-us.ahrq.gov/toolssoftware/comorbidity/comorbidity.jsp](https://www.hcup-us.ahrq.gov/toolssoftware/comorbidity/comorbidity.jsp)
- Initial Patient Assessment information (completed on the day of admission)
  - Living situation (alone, family, group setting)
  - Indicators of autonomy level (ability to use bell, autonomy in activities of daily life, difficulty ambulating or swallowing)

A.2 Dynamic variables

The dynamic variables we considered are:

- From general admission data
  - Number of days the patient already spent at the hospital during this admission,
  - Daily number of admission/discharges at a hospital/ward level,
  - Daily number of (long-stay) patients at a hospital/ward level,
  - Current day of the week,
  - Current service the patient is in. This variable is used to identify off-service patients,
– Indicator of whether the patient is currently in an ICU,
– Number of days spent in the ICU so far,
– Number of days spent with the same attending,

• From vital sign measurements
  – We monitored heart rate, respiratory rate, temperature, O2 saturation, diastolic/systolic blood pressure, pain level, activity level, RASS score.
  – For those vital signs, we reported last value at the end of the day, average daily value and trend based on past 4 measurements.

• From lab result information
  – We only included blood results for they are widely performed for all inpatients.
  – We counted the number of lab measurements at an abnormal value.
  – For platelet count, white cells count, hematocrit, chloride, sodium, potassium, RDW, RDW SD and urea nitrogen, we reported the last value, the last slack compared to the normal range and the slope based on previous measurements.
  – For bilirubin, glucose, PTT, sedimentation rate, troponin, albumin and INR, we created a discrete variable taking values 0/1/2: 0 if the quantity was never measured for the patient; 1 if the quantity was measured and currently at a normal level; 2 if the quantity was measured and currently at an abnormal level.

• From medical orders
  – We divided orders into types: Blood bank, Cardiology, Consultation, Critical Care, General Care, Lab, Neurology, Nutrition, Obstetrics, Radiology, Respiratory, IV. Observe that each order can be a new order or a change/termination of a previously placed order.
  – For all order types, we daily counted the number of orders placed (counting +1 for new and -1 for discontinuing orders) and the number of order changes.
  – For some types, namely IV, TPN, Blood bank, we performed those counting operations at a sub-type level as well.

• From pharmacy data
  – We grouped medications based on the AHFS Pharmacologic-Therapeutic Classification (first two digits of the AHFS code).
  – Within each group, we computed the number of medications the patient was currently taking, the number of medications the patient started/stopped taking on each day and the time (in days) she had been taking these medications for.

• For diagnosis codes
  – We used the number of ICD10 codes reported since admission as a measure of a patients severity.
  – We group ICD10 codes based on their letter and 1) computed the proportion of codes within each category to identify the type of illness, 2) computed the gini score of the distribution of code letters to measure clinical complexity.

• From the OR schedule
  – Future surgery: we computed and reported the inverse of the time-to-next-surgery. This quantity equals 0 if no surgery is scheduled.
  – Past surgeries: we counted the surgeries the patient had since admission (by number and duration). We reported the inverse of the time-to-last-surgery as well.

• From documentation and notes
  – Without using raw information from the documentations or notes about the patient, we recorded the number of documents/notes entered about the patient on each day, to measure the level of activity around each patient.
  – We also computed the inter-day difference to capture variation in activity.