## Research and Applications

# Predicting next-day discharge via electronic health record access logs

Xinmeng Zhang[1], Chao Yan [ID][1], Bradley A. Malin[1,2,3], Mayur B. Patel[4,5], and
You Chen [ID][1,2]

[1]Department of Computer Science, Vanderbilt University, Nashville, Tennessee, USA, [2]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA, [3]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, USA, [4]Section of Surgical Sciences, Departments of Surgery & Neurosurgery, Division of Trauma, Surgical Critical Care, and Emergency General Surgery, Nashville, Tennessee, USA, and [5]Geriatric Research and Education Clinical Center, Surgical Services, Veteran Affairs Tennessee Valley Healthcare System, Nashville, Tennessee, USA

Xinmeng Zhang and Chao Yan are first authors.

Corresponding Author: You Chen, PhD, Department of Biomedical Informatics, Vanderbilt University Medical Center, Suite 1475, 2525 West End Avenue, Nashville, TN 37203, USA; you.chen@vumc.org

## ABSTRACT

**Objective:** Hospital capacity management depends on accurate real-time estimates of hospital-wide discharges. Estimation by a clinician requires an excessively large amount of effort and, even when attempted, accuracy in forecasting next-day patient-level discharge is poor. This study aims to support next-day discharge predictions with machine learning by incorporating electronic health record (EHR) audit log data, a resource that captures EHR users' granular interactions with patients' records by communicating various semantics and has been neglected in outcome predictions.

**Materials and Methods:** This study focused on the EHR data for all adults admitted to Vanderbilt University Medical Center in 2019. We learned multiple advanced models to assess the value that EHR audit log data adds to the daily prediction of discharge likelihood within 24 h and to compare different representation strategies. We applied Shapley additive explanations to identify the most influential types of user-EHR interactions for discharge prediction.

**Results:** The data include 26 283 inpatient stays, 133 398 patient-day observations, and 819 types of user-EHR interactions. The model using the count of each type of interaction in the recent 24 h and other commonly used features, including demographics and admission diagnoses, achieved the highest area under the receiver operating characteristics (AUROC) curve of 0.921 (95% CI: 0.919–0.923). By contrast, the model lacking user-EHR interactions achieved a worse AUROC of 0.862 (0.860–0.865). In addition, 10 of the 20 (50%) most influential factors were user-EHR interaction features.

**Conclusion:** EHR audit log data contain rich information such that it can improve hospital-wide discharge predictions.

**Key words:** hospital capacity management, daily discharge prediction, EHR audit logs, user-EHR interactions

## INTRODUCTION

Poor operation of hospital capacity management often prolongs a patient's waiting time until admission (ie, delayed care) and discharge (ie, prolonged length of stay). Both of these delays lead to inefficient planning and utilization of care resources, increased healthcare costs, and consequently, a higher risk of in-hospital complications and mortality.[1,2] Uncertainty in when patients will be discharged from the hospital is a prominent contributor to capacity management problems in that it challenges care teams in delivering appropriate care and can drive the inpatient flow into undesirable scenarios, such as overcrowding.[3] As a consequence, hospital administrators have paid an increasing amount of attention to the discharge prediction problem to align timely care services with a patient's needs and streamline inpatient flow of hospitals.[1,4–11]

Prediction of which patient will be discharged and when is typically based on a clinicians' assessments, whereby predictions are made by manually reviewing a patient's clinical status, as well as their anticipated treatment.[12,13] This process begins at a patient's admission and is updated throughout their stay.[14,15] Notably, real-time demand capacity management, which was designed by the Institute for Healthcare Improvement and has been implemented at multiple hospitals, has shown promise in improving inpatient flow via a set of assessments centered on the same-day discharge prediction made in morning clinician huddles.[16] However, such clinical assessments require a nontrivial amount of hospital-wide manual effort, which may reduce clinician time spent on patient care and increase their workload.[7,17] Moreover, clinicians can bias the prediction subjectively due to personal experiences and training levels.

A wide variety of machine learning models have been developed to automate discharge prediction based on data from electronic health records (EHRs).[17–25] For instance, Barnes et al[17] learned a tree-based model to predict (at 7 AM of each day) the same-day discharge before 2 PM and 12 PM. This model was significantly better than physicians' in the general inpatient setting. Safavi et al[26] developed a neural network to predict the discharge likelihood of a surgical care patient within 24 h, outperforming a clinician surrogate model by a wide margin. Levin et al[19] evaluated multiple clinical unit-specific models that relied on real-time EHR data to assess same, as well as next, day discharge and demonstrated their effectiveness in reducing the length of stay.

Though EHR-derived features have been integrated into these models, they are mostly those that directly indicate a patient's health status, such as diagnoses, procedures, laboratory tests, and medications. A critical source of information that has been neglected is EHR audit logs, which capture EHR users' interactions with patient records and provide clues into the complexity of care delivered.[27–33] This type of data is widely collected by EHR systems and communicates various semantics (eg, orders, messages, flowsheets, chart review notes, and general documentation), such that they have been utilized to support many endeavors, including health process modeling,[34–36] clinical task and workflow analysis,[29,37–42] time-motion study of physicians' work in EHRs,[29,38–41,43] user-user interaction structure analysis,[29,38–41] and privacy audits.[44,45] More importantly, this data source has the potential to reflect clinicians' assessment and insights into a patients' clinical status in the decision-making process. For example, frequent scanning of medical device barcodes may indicate that the patient requires a significant amount of attention and, thus, is not ready for discharge. Similarly, a high frequency in nurse monitoring activities may indicate that the patient's status is uncertain. Though there is certainly a correlation in the informa-

tion documented in the EHR and the audit logs, we hypothesized that the latter can provide new signals informing patients' discharge status. This hypothesis stems from the fact that EHR audit logs associated with a patient may reflect the insights from all of the members of their care team, which should intuitively provide a broader picture of patient care than previous discharge prediction models.

In this study, we incorporate user-EHR interactions, as documented in audit log data, to develop machine learning models for discharge prediction. We focus on the next-day discharge task, which can provide the capacity management team with sufficient response time to plan and has proven to be a challenge for physicians.[12] We show that incorporating such data can significantly improve the performance of discharge prediction over time. We evaluate the prediction model in the general inpatient setting of Vanderbilt University Medical Center (VUMC) and identify the most influential risk factors in prediction.

## MATERIALS AND METHODS

### Dataset and outcome definition

This study was conducted at VUMC, which is an 1162-bed teaching hospital using the EPIC EHR system . We extracted all adult (≥18 years old) inpatient visits in 2019. We included all visits with a discharge of home or care facilities and ruled out visits for patients with a stay shorter than 24 h or who died during hospitalization. For patients with more than one inpatient, we retained the first one to mitigate bias in our analysis. The resulting dataset contained data on 26 283 patients with an average age of 52.9 (±19.5) and a male to female ratio of 45:55.

For each visit, we collected the following information from the patient's EHR: (1) demographics (in the form of age, gender, and self-reported race), (2) historical diagnoses from 2005 up to admission, (3) admission diagnoses, (4) the time period of the visit, (5) category of enrolled insurance program (in terms of public, private, or uninsured), and (6) body mass index (BMI) and heart rate (HR) readings during the visit. We mapped all historical ICD-9/10 diagnosis codes to Phenome-wide association study codes (Phecodes), which aggregate billing codes into clinically meaningful phenotypes to reduce sparsity in the data.[46] We used the clinical classifications software categorizations of admission diagnoses as a proxy for the reason for a visit. Table 1 summarizes the characteristics of the final dataset.

We extracted timestamped user-EHR interactions for each visit from the EHR audit logs. There were 819 distinct user-EHR interaction types, which fall into 4 high-level event groups: (1) view (eg, an EHR user accessed a treatment plan), (2) modify (eg, a resident added new content into a progress note), (3) export (eg, an operating room report was printed by a nurse), and (4) system (eg, an e-consult message was sent). Although there are no widely adopted standards for defining the types (regarding the semantics, granularity, and aggregation levels) of user-EHR interactions in EHR systems, such types are typically based on commonly understood categories of EHR tasks.[27] Determining an interaction type usually involves (1) the specific function of the EHR system that is utilized and (2) the role of the EHR user (eg, physician or nurse) who performs this function. The second aspect is explicitly communicated when there is a binding between a specific role and an EHR function (eg, "Patient notes loaded for review by clinicians"). Researchers and EHR vendors (eg, Epic[28] and Cerner[47]) have created lower-level common event categories, such as chart review imaging, laborato-

**Table 1.** Summary of the dataset used in this study

| Characteristic | Distribution | |
|---|---|---|
| Age at admission | 35.2, 54.3, 68.6 | 52.9 ± 19.5 |
| Race | | |
|   White | 78.1% | 20 536 |
|   Black | 15.5% | 4066 |
|   Asian | 1.8% | 473 |
|   American Indian or Alaska Native | 0.17% | 45 |
|   Others | 4.42% | 1163 |
| Gender | | |
|   Male | 45.0% | 11 838 |
|   Female | 55.0% | 14 445 |
| Insurance type | | |
|   Commercial | 42.1% | 11 074 |
|   Medicare/Medicaid | 51.2% | 13 460 |
|   None | 6.7% | 1749 |
| User-EHR interactions | | |
|   Total number of interactions before discharge | 118, 215, 457 | 331.0 ± 313.5 |
|   Number of unique interactions before discharge | 32, 49, 82 | 59.5 ± 38.7 |
| Unique Phecodes before admission | 7, 18, 39 | 30.0 ± 35.6 |
| Most frequent discharge units | | |
|   Postpartum | 12.6% | 3324 |
|   Cardiac stepdown | 6.4% | 1675 |
|   Neuroscience | 4.8% | 1259 |
|   Transplant and surgery | 4.6% | 1212 |
|   Spine | 3.9% | 1016 |
|   Urology | 3.8% | 998 |
|   Medicine cardiac stepdown | 3.7% | 977 |
|   Surgical stepdown | 3.7% | 973 |
| Length of stay (days) | 2.3, 3.7, 6.0 | 5.6 ± 6.4 |
| Discharge time | | |
|   8 PM—8 AM | 2.5% | 665 |
|   8 AM—2 PM | 52.5% | 13 796 |
|   2 PM—8 PM | 45.0% | 11 822 |
| Discharge day of week | | |
|   Monday | 14.3% | 3766 |
|   Tuesday | 15.5% | 4081 |
|   Wednesday | 15.8% | 4163 |
|   Thursday | 15.6% | 4108 |
|   Friday | 18.0% | 4728 |
|   Saturday | 11.5% | 3025 |
|   Sunday | 9.2% | 2412 |

*Notes* : $a, b,$ and $c$ represent the first quartile, median, and third quartile. $x \pm y$ represents the mean and 1 standard deviation. $x\%$ $y$ represents that the percentage of $y$ patients (in a given category) is $x\%$ among all patients.

EHR: electronic health record.

ries, medications, notes, messaging, order entry, billing and coding, and inbox (refilling medications). In this study, we used the Epic user-EHR interaction categories (fine-grained level) in the prediction model. Note that the extracted use-EHR interactions do not include the identities of healthcare workers or patients in their definitions.

The primary outcome of this study is 24-h discharge status (ie, whether each patient would be discharged within 24 h). We followed the prediction setting in Ref.[17] and chose 2:00 PM as the prediction time point, a time that can (1) help mitigate the overcrowding typically seen in the late afternoon and evening and (2) reasonably inform planning for new admissions in the following day.[16]

## Model development

To train and evaluate next-day discharge prediction models, we partitioned each visit on a daily basis and built an instance for each day—except for the last day. Each instance contains information up to 2:00 PM on the prediction day. This produced 133 398 instances. We then paired each instance with a binary label indicating the discharge status by 2:00 PM of the next day. This led to a positive-negative ratio of approximately 1:4.

We primarily relied upon the light gradient boosting machine (LGBM) framework,[48] a light implementation of gradient-boosted ensembles of decision trees, to learn predictive models. In comparison to traditional classifiers (eg, logistic regression or support vector machine), LGBM has outperformed in prediction tasks and provides interpretable predictions.[49,50] We developed multiple LGBM models to investigate the impact of EHR audit log data to the performance of next-day discharge prediction.

The complete set of features used in this study, as well as the input strategy for each model, is shown in Figure 1. We incorporated 2 types of features in model development. Static (baseline) features, whose values remain constant across a visit, include age (<20, 20–

**Figure 1.** Input feature space for the LGBM models. Each number in the gray row (bottom) indicates the number of values associated with a feature type (eg, there are 9 age groups of equal width). LGBM: light gradient boosting machine.

29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, or $\geq$90), gender, race (White, Black, Asian, Alaska and Native American, or Others), insurance type (Medicaid/Medicare, commercial, or no insurance), historical diagnoses, and diagnoses present at admission. By contrast, daily features, the values of which dynamically change during a visit, include the nearest HR and BMI readings to the prediction time point, user-EHR interactions, the number of days from admission, and day of week. Note that HR and BMI are represented as the actual values and an indicator for missingness. Each instance is a concatenation of static features and daily features.

We learned 5 types of models described below. To name these models, we used "N," "B," and "C" to represent the numerical, binary, and cumulative representation strategies of user-EHR interactions, respectively.

- *LGBM-N-INT*: Each dimension of the user-EHR interaction domain indicates the numerical counts of the corresponding interaction type that occurred during the past 24 h.
- *LGBM-B-INT*: Binary indicators are used to represent the presence (or absence) of user-EHR interaction types during the past 24 h.
- *LGBM-C-INT*: Instead of considering the user-EHR interactions for only 24 h, this model uses the counts of each user-EHR interaction type accumulated from admission.
- *LGBM-NO-INT*: This model uses all features, except user-EHR interactions.
- *LGBM-ONLY-INT*: This model takes only the numerical counts of each user-EHR interaction type during the past 24 h.

The first 3 models are designed to investigate the impact of different representation strategies for user-EHR interaction data on the prediction problem, while the latter 2 models investigate the impact of user-EHR interaction more generally.

Given that it is challenging for the LGBM framework to embed the temporal relationship in user-EHR interactions and historical diagnoses, we built a bidirectional RETAIN-based model,[51] an advanced variation of a recurrent neural network, as an alternative model. The architecture of this model is shown in Supplementary Appendix Figure SA.1. In addition, we compared our best-performing model to a random forest-based discharge prediction tool developed by Barnes et al.[17] We also compared the LGBM architecture to the random forest by training another LGBM model, which uses the same set of features as the random forest-based model.

## Model training, evaluation, and calibration

We randomly split the inpatient visits into a training and validation set (85%), a calibration set (5%), and a test set (10%). This process was repeated 5 times to enhance the robustness of the investigation. In each split, we performed 5-fold cross-validation and ensured that the data for each patient appears in one set only. We used Youden's Index,[52] which equally weights false positives and false negatives, to determine the thresholds for positive or negative prediction. We use standard performance evaluation measures, including the area under the receiver operating characteristic (AUROC) curve, the area under the precision-recall curve (AUPRC), accuracy, recall (sensitivity), specificity, positive predictive value (PPV), and negative predictive value. To assess whether the performance improvement achieved through EHR audit log data is statistically significant, we applied a 2-tailed paired $t$ test ($n = 25$). A grid search technique was applied to tune the hyperparameters for the LGBM models (the details of which are provided in Supplementary Appendix SB). In addition to the overall performance, we evaluated how the best model performs in different subpopulations and scenarios, by varying age, length of stay at the prediction point (ie, elapsed days since admission), and care unit.

Before evaluating models on the independent test set, we applied isotonic regression[53] on the calibration set to perform model calibration.[54] Expected calibration error (ECE) was used to assess how well a model was calibrated.[53] To compute ECE, model predictions (a surrogate probability of being positive) were sorted and then divided into $M$ bins with equal sizes. We then measured the weighted absolute differences between models' accuracy (ie, the ratio of correct predictions in a given bin) and confidence (ie, the average probability in a given bin) across all bins. Formally, ECE is defined as:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \text{acc}(B_m) - \text{conf}(B_m) \right|,$$

where $n$ and $B_m$ denote the total number of predictions and the set of predictions with confidence belonging to the $m$th bin, respectively. To visualize the 2 quantities in each bin, we drew reliability curves[49] by setting $M = 20$.

## Influential risk factors

In addition to assessing prediction performance, we investigated the prediction interpretability. We calculated the SHapley Additive

exPlanations (SHAP) values for each feature[55] in the best-performing model. Specifically, we measured the mean SHAP and the standard deviation for each feature in all data splits. We then identified influential risk factors by ranking features based on their SHAP values.

## RESULTS

### Overall performance

Figure 2 depicts the performance for all prediction models. There are 2 notable findings regarding the representation strategies of user-EHR interactions and the impact of these data to the next-day discharge prediction. First, LGBM-N-INT achieved the highest AUROC (mean = 0.921, 95% confidence interval [CI]: 0.919–0.923), corresponding to an AUPRC of 0.820 (95% CI: 0.818–0.823). By outperforming LGBM-B-INT (*P* value: $6.20 \times 10^{-17}$) and LGBM-C-INT (*P* value: $3.01 \times 10^{-18}$), it appears that the number of user-EHR interactions in the past 24 h is more informative than its binary representation (ie, any interaction or not) or the cumulative number of interactions since a patient's admission. Second, when the model is composed solely of user-EHR interactions (ie, LGBM-ONLY-INT), it can achieve a very high AUROC of 0.890 (95% CI: 0.888–0.892), which is slightly lower than LGBM-B-INT. However, by contrast, the model that focused solely on the more traditional features and lacked user-EHR interactions (ie, LGBM-NO-INT), led to an obvious decline on both AUROC (0.862) and AUPRC (0.662). Such a phenomenon suggests that user-EHR interactions in the past 24 h are driving the prediction performance. In addition, RNN-BASED achieved a worse AUROC of 0.912 (95% CI: 0.910–0.913) and a similar AUPRC of 0.820 (95% CI: 0.818–0.822) comparing to LGBM-N-INT, indicating that the design of capturing temporal relationships in user-EHR interaction sequences might not improve the prediction performance.

Table 2 summarizes the results of all models on measures except for AUROC and AUPRC. Notably, LGBM-N-INT successfully predicted 77.8% of the actual next-day discharges with an accuracy of 87.7%. We also observed 29.4% (46.8%) of the false positives corresponded to a discharge within 30 (48) h from the prediction time point. These 2 endpoints correspond to 8 PM of the next day and 2 PM of the day after next day. Supplementary Appendix SC provides a detailed analysis of the impact of length of stay on the prediction point. In comparison to the random forest-based discharge prediction tool based on Barnes et al,[17] LGBM-N-INT demonstrates better-discriminating capability in terms of AUROC. Meanwhile, the random forest-based model performs significantly worse than its LGBM variant. Detailed results are provided in Supplementary Appendix SD.

After calibrating model output (via isotonic regression), we obtained sufficiently small ECE scores for all models, corresponding to
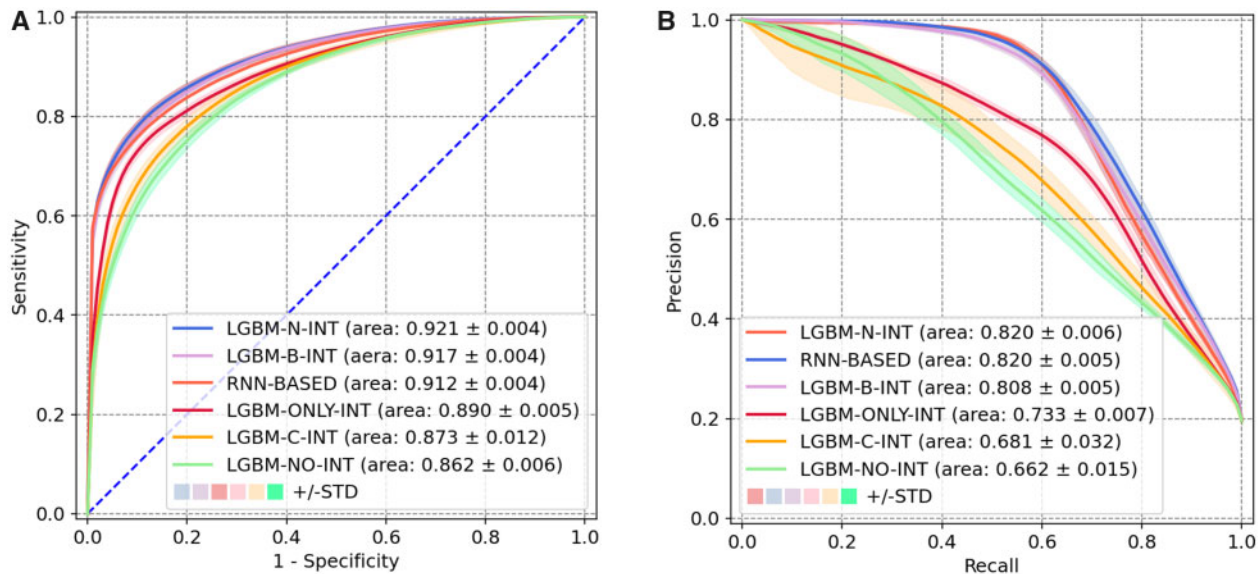


**Figure 2.** Next-day discharge prediction model performance. (A) Receiver operating characteristic curve. (B). Precision-recall curve.

**Table 2.** Performance indicators for next-day discharge prediction models

| Model | Accuracy | Recall | Specificity | PPV | NPV |
|---|---|---|---|---|---|
| **LGBM-N-INT** | **0.877 (0.872–0.883)** | **0.778 (0.769–0.786)** | **0.901 (0.893–0.909)** | **0.654 (0.639–0.669)** | **0.944 (0.942–0.946)** |
| **LGBM-B-INT** | 0.868 (0.865–0.872) | 0.778 (0.771–0.785) | 0.890 (0.884–0.895) | 0.628 (0.612–0.644) | 0.944 (0.942–0.945) |
| **RNN-BASED** | 0.868 (0.860–0.875) | 0.763 (0.751–0.776) | 0.893 (0.881–0.905) | 0.637 (0.614–0.661) | 0.940 (0.938–0.943) |
| **LGBM-ONLY-INT** | 0.855 (0.847–0.864) | 0.750 (0.737–0.763) | 0.880 (0.868–0.893) | 0.604 (0.563–0.646) | 0.936 (0.934–0.939) |
| **LGBM-C-INT** | 0.820 (0.812–0.828) | 0.744 (0.734–0.753) | 0.838 (0.827–0.849) | 0.525 (0.490–0.561) | 0.932 (0.930–0.934) |
| **LGBM-NO-INT** | 0.805 (0.798–0.812) | 0.727 (0.717–0.737) | 0.823 (0.813–0.833) | 0.498 (0.468–0.527) | 0.926 (0.924–0.929) |

*Notes:* The models are sorted by AUROC. *x* (*a*, *b*) represents mean value and 95% confidence interval. The highest score for each metric is in bold.

AUROC: area under the receiver operating characteristics; LGBM: light gradient boosting machine; NPV: negative predictive value; PPV: positive predictive value.

a neglectable gap between the predicted next-day discharge probabilities and the actual proportion of discharges in each probability bin. Supplementary Appendix SE provides detailed results for the model calibrations. Figure 3 shows the reliability curves of LGBM-N-INT for the pre- and postcalibration scenarios. The 2 curves are well aligned with the ideal (diagonal) line, and it is clear that LGBM-N-INT is well-calibrated even without the *post hoc* calibration, as indicated by an ECE score of 0.014 (95% CI: 0.013–0.016) that is almost equivalent to its postcalibrated score of 0.011 (95% CI: 0.009–0.012).

## Evaluations on patient subpopulations

We applied LGBM-N-INT to different subpopulations with varying age, prediction day, and care area, and evaluated its performance accordingly.

### Age

The AUROC and AUPRC of LGBM-N-INT are not uniformly distributed across all age groups. Figure 4A and B shows that patients in age the 40–50 age group have the highest AUROC (0.930) and AUPRC (0.839) with a recall of 0.795 (Figure 4C) and a precision (or PPV) of 0.688 (Figure 4D). From the perspective of variability from mean values, both AUROC and AUPRC demonstrate a larger standard deviation on patients younger than 20 years old. This may be explained by the small size of this group (accounting for 1.53% of the total number of patients). In addition, the average AUROC and AUPRC in this age group were the lowest (0.879 and 0.764) among all age groups.

The gray curve in Figure 4C shows the number of positive instances for each age group in a test set. Figure 4A and C illustrates that for patients 40 years old and younger, there is a direct correlation between group size and score for AUROC and AUPRC. However, this relationship inverts for patients over 40 years old. This suggests that the next-day discharge prediction for elderly patients is more challenging than younger adults even with a larger number of instances. It appears that this is because elderly patients, in general, have a health condition that is more complex than younger adults, leading to more uncertainty in management.

### Prediction day

Figure 5 shows the model performance by varying day of making predictions, where we aligned the admission day of all inpatients. Predictions started from 2 PM on the second day of a visit. We grouped together all instances corresponding to a stay longer than 1 week. The highest AUROC (0.931 in Figure 5A) and AUPRC (0.847 in Figure 5B) appear on different days. Nevertheless, both AUROC and AUPRC achieved their lowest values (0.888 and 0.734) at day 2, which is likely due, in part, to an insufficient amount of information available. The number of actual discharge events (equivalent to the number of patients discharged before 2 PM of the next day) changes significantly with day of making predictions (Figure 5C). Specifically, the number of actual discharge events peaks at day 3 and decreases monotonically through day 6. However, despite the quick decline in the number of actual discharges, the AUROC and AUPRC remain relatively stable.

### Care area

Inpatients, especially those with complex conditions, can be moved and treated at more than one care units (or locations) while in a hospital. For example, it is common for inpatients to begin in the emergency department and then move to a specific unit of care. It is also common for a patient to move from an intensive care unit to an operating room. Given the complexity of binning patients into care units, we show the performance breakdown of LGBM-N-INT based on areas of VUMC that were finally visited before discharge in Table 3. Here, a patient care area contains multiple care units, which can be found in Supplementary Appendix Table SAF.1.

The next-day discharge predictions for adult inpatients, whose final care areas before discharge were Acute Care Medicine, Acute Care Surgical or Women's Health, demonstrate a higher AUPRC than the overall AUPRC (0.820). Notably, predictions for patients who were last at Adult Oncology show higher performance on both metrics than average. Supplementary Appendix Table SAF.2 provides a more detailed presentation of these results.



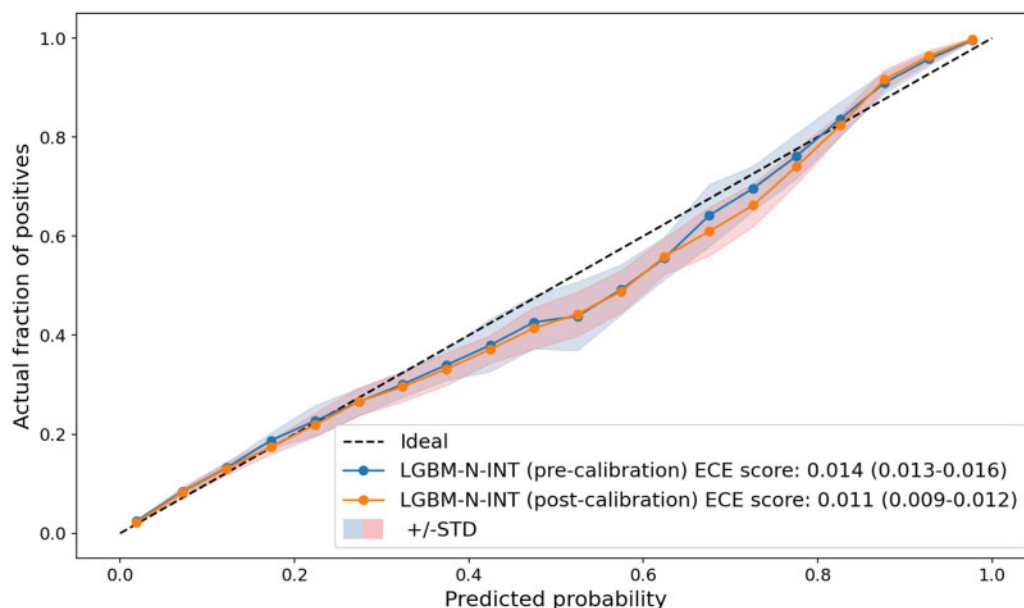**Figure 3.** Reliability curve for the LGBM-N-INT model on the test sets. LGBM: light gradient boosting machine.
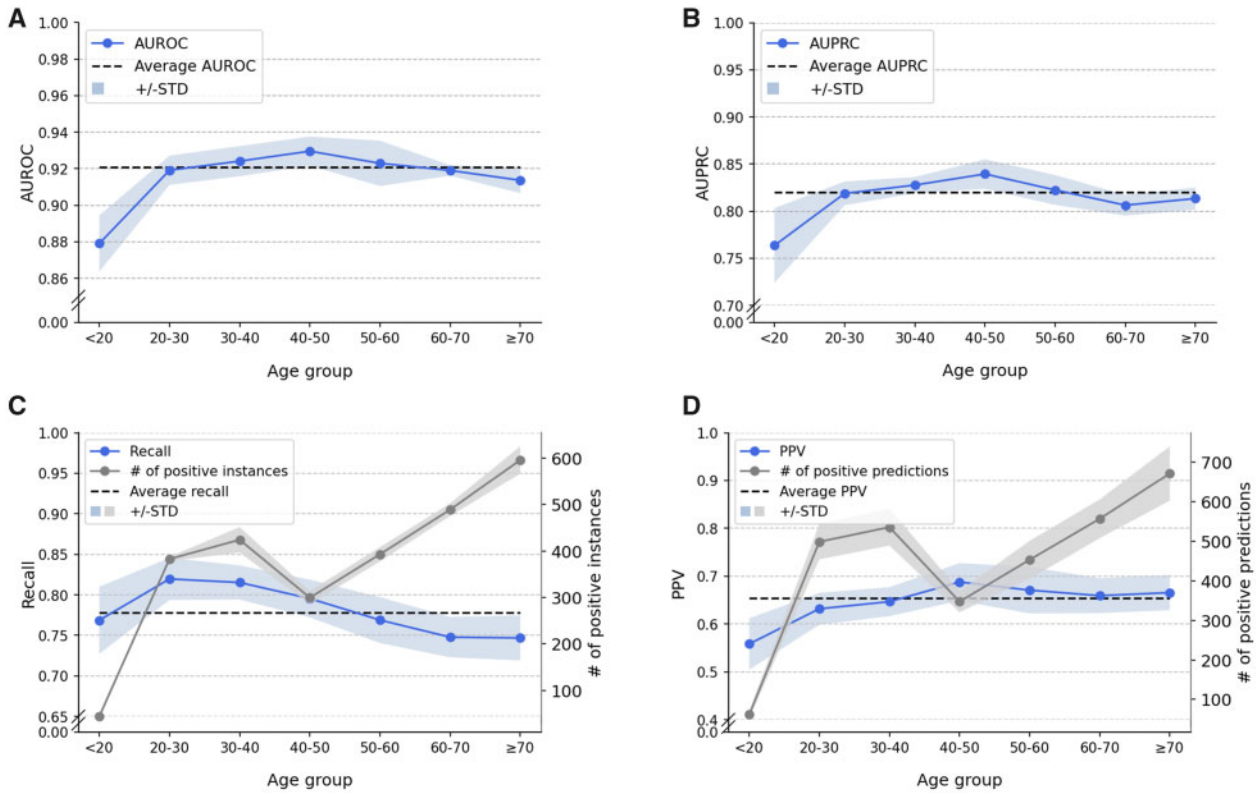
**Figure 4.** Influence of age on LGBM-N-INT next-day discharge prediction: (A) AUROC, (B) AUPRC, (C) Recall, and (D) PPV. AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristics; LGBM: light gradient boosting machine; PPV: positive predictive value.



**Figure 5.** Influence of prediction day on LGBM-N-INT next-day discharge prediction: (A) AUROC, (B) AUPRC, (C) Recall, and (D) PPV. AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristics; LGBM: light gradient boosting machine; PPV: positive predictive value.
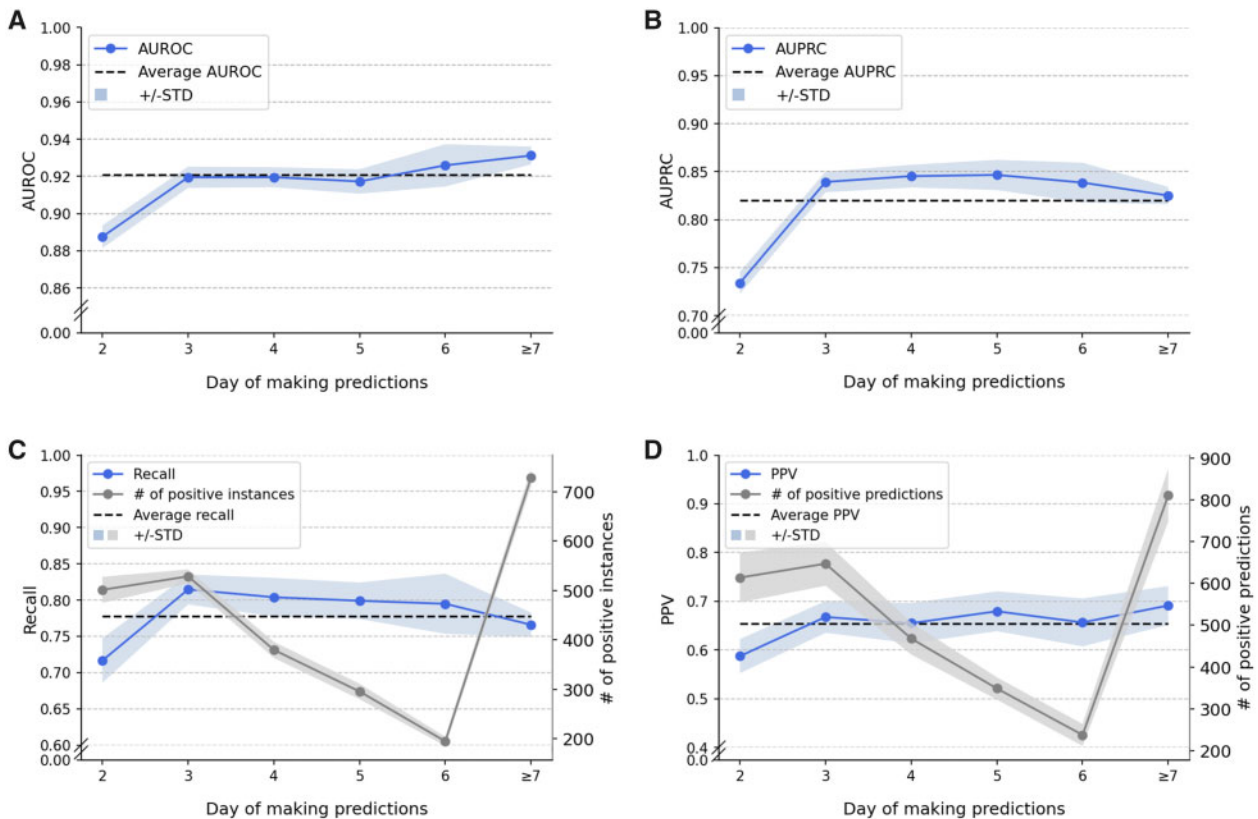
**Table 3.** Performance breakdown of LGBM-N-INT based on finally visited care areas before discharge

| Care area | AUROC | AUPRC |
|---|---|---|
| Overall | 0.921 (0.919–0.923) | 0.820 (0.818–0.823) |
| Cardiac stepdown | 0.923 (0.920–0.925) | 0.801 (0.796–0.806) |
| Acute care medicine | 0.917 (0.915–0.920) | **0.823 (0.819–0.826)** |
| Acute care surgical | 0.919 (0.915–0.923) | **0.825 (0.821–0.829)** |
| Critical care | 0.922 (0.920–0.925) | 0.802 (0.796–0.807) |
| Oncology | **0.935 (0.932–0.938)** | **0.836 (0.831–0.842)** |
| Women's health | 0.896 (0.887–0.904) | **0.853 (0.846–0.859)** |

*Notes:* Values that are significantly higher than the overall performance (shown in the first row) of LGBM-N-INT are highlighted in bold. $x$ ($a$, $b$) represents mean value and 95% confidence interval.

AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristics; LGBM: light gradient boosting machine.

## Most influential factors

We performed a feature analysis on LGBM-N-INT since it was the best-performing model. Figure 6A shows the 20 most influential factors in terms of SHAP values. To indicate the category of each factor, we use [INT], [LOS], [Historical Diagnosis], [Weekday], [Insurance], and [Admission Diagnosis] as prefix to denote user-EHR interactions, length of stay (in days) at prediction point, historical diagnoses before admission, day of week, insurance types, and diagnoses on the day of admission, respectively. Ten (or 50%) of these factors correspond to user-EHR interactions with distinct semantics. In particular, the number of barcode scan events, Nurse's station mode monitoring (which is one of the EHR system monitoring functions, and the information relevant to high-acuity patients, such as infection status and deterioration scores, will be shown when this function is utilized) events, clinical note view events, inpatient medication administration record access events, and result review access events in the past 24 h are the top 5 factors. For each of them, the higher the value, the more this factor contributes to a nondischarge prediction.

In Figure 6B and C, we provide 2 examples to demonstrate the correlation between the specific values of a given feature and the corresponding SHAP values. We observed in Figure 6B that when there are less than 16 barcode scan events in the past 24 h, this factor contributes positively to a next-day discharge prediction; however, it contributes to a prediction of nondischarge (ie, remain in hospital) for the numbers greater than 23. The contribution of this factor demonstrates a monotonically decreasing trend with the increase of the number of barcode scans. From Figure 6C, we can see that the presence of Nurse's station mode monitoring access tends to contribute to a nondischarge prediction. A dense record of these 2 events suggests that a patient might be in the midst of treatment or has been unstable status during the past 24 h.
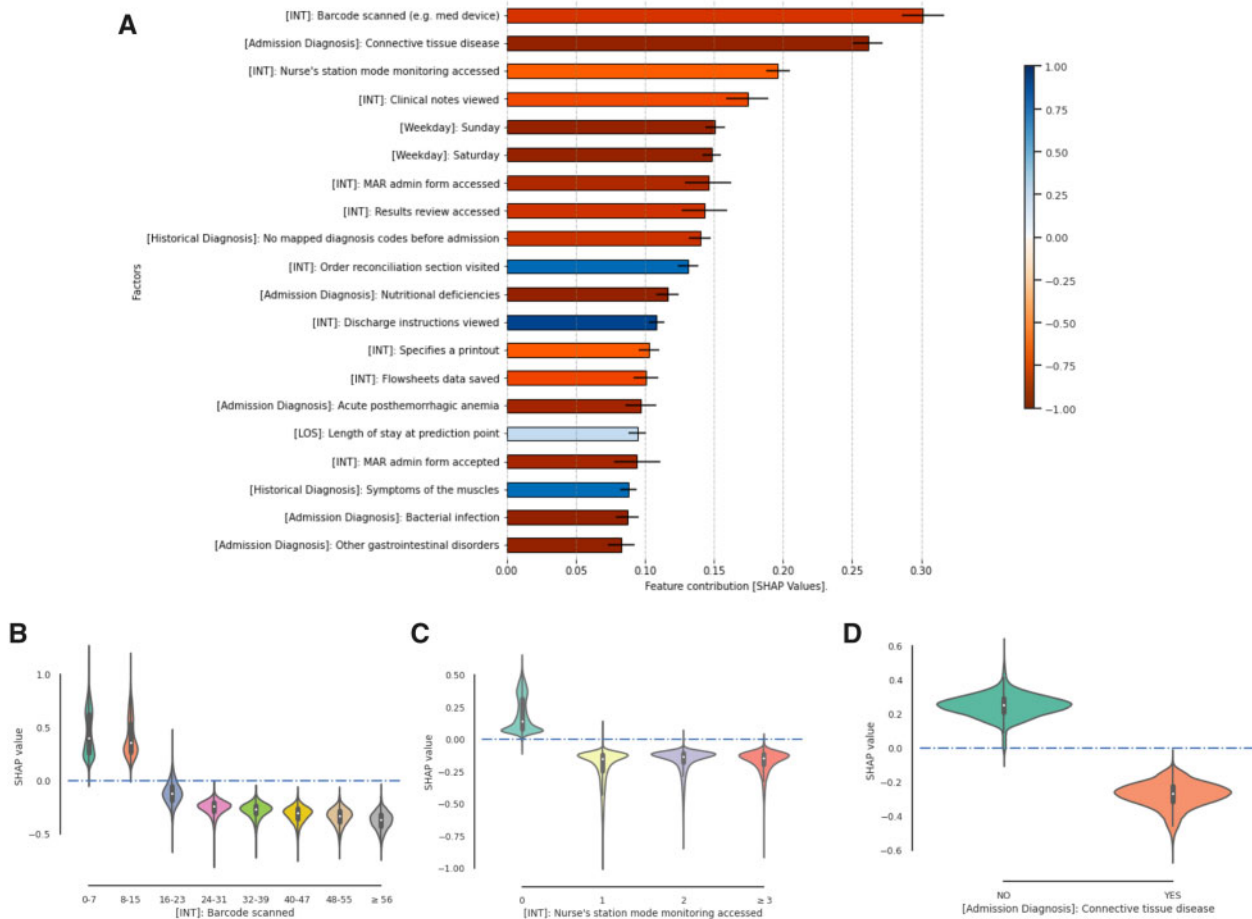


**Figure 6.** Feature importance in the next-day discharge prediction model. (A) Top risk factors indicated by SHAP values. The color of each bar indicates the Pearson correlation coefficient between SHAP values and values of the corresponding factor. (B–D) Violin plots of factor values against the corresponding SHAP values on 3 factors: [INT] Barcode scanned, [INT] Nurse's station mode monitoring accessed, and [Admission Diagnosis]: Connective tissue disease. MAR: medication administration record; SHAP: SHapley Additive exPlanations.

The 2 user-EHR interaction features that are positively correlated with their SHAP values are (1) visiting the order reconciliation section and (2) viewing discharge instructions. Order reconciliation is an important process in the EHR discharge navigator, which aims to help physicians reconcile problems and medication list for discharge. For example, physicians are supposed to indicate the set of resolved and unresolved problems (for follow-ups after discharge) and to determine whether to modify, prescribe, and stop prescribing certain drugs for discharge. Viewing discharge instructions seems a straightforward signal to discharge. We investigated the numbers of actual discharges, predicted discharges, and their intersections conditioned on the presence/absence of these 2 types of interactions in the past 24 h (Supplementary Appendix Figure SG.1), respectively. We found that only 16.6% of the instances with the presence of order reconciliation section visits, and 16.3% of the instances with the presence of discharge instruction views corresponded to a discharge event in the next 24 h; by contrast, there are still 20.7% and 20.1% of instances without the presence of the 2 user-EHR interactions corresponding to a discharge event in the next 24 h, respectively.

In addition to the user-EHR interactions, several other features also contribute to the prediction. The second most influential factor is an admission diagnosis—connective tissue diseases. We observed in Figure 6D that the presence of this diagnosis category contributes to a nondischarge prediction. The absence of historical diagnosis also tends to contribute to a nondischarge prediction. Specifically, the presence of muscle symptoms gives an obvious positive impact to discharges compared to other preadmission diagnoses. Sunday and Saturday are also among the predictors of nondischarge predictions of the next day.

## DISCUSSION AND CONCLUSION

Efficient capacity management of hospitals, which sets the stage for timely and appropriate patient care, demands the minimization of uncertainty in patient flow, and, as a consequence, requires an accurate estimation of discharges in real time. This study is designed to provide hospital administrators an analysis tool to predict hospital-wide discharges 1 day ahead. Our research is the first to incorporate EHR audit logs (more precisely, user-EHR interactions) to predict discharges. In this study, we verified that such data can substantially contribute to the prediction of next-day discharges. Notably, our findings indicate that the count of each type of user-EHR interaction in the past 24 h appears to be the most effective representation for such information.

EHR audit log data have been used to investigate time-motion activities of EHR users with respect to EHR systems.[28,29,38,41,43] Various studies have shown that they are useful for measuring the amount of time a clinician spends in the EHR system,[28,56] which was relied upon to measure EHR-related workload, burden, and burnout.[30,32,33] In addition, EHR audit log data analysis can identify collaborative patterns between EHR users through user-EHR interactions.[57,58] Based on the aforementioned studies on EHR audit log data, this study assumed time-motion activities of a collective of users who comanage a patient could be leveraged to indicate the patient's discharge status. Notably, our research demonstrates that the activities of a collective of EHR users are markers of a patient's discharge status.

EHR audit log data are indicative of the discharge status of an inpatient in several ways. First, EHR audit logs communicate granular semantics that describe the complex healthcare workflow of an inpatient, including signals regarding the stage the patient is in with respect to hospital treatment and services. For example, the co-occurrence of "Barcode scanned (eg, med device)" and "Nurse's station model monitoring accessed" within the past 24 h is more likely a sign of active treatment of an unstable patient, such that patients with this pattern may require a longer period hospitalization. Quantitatively, this pattern occurs in 38 241 daily instances of patients, whereas only 8.7% correspond to a discharge event within the next 24 h. Second, the frequency of distinct user-EHR interaction types provides additional signals for inferring a patient's discharge status. The discharge percentage decreases to 6.3% when requiring more than 24 occurrences of "Barcode scanned (eg, medical device)" in the past 24 h. Third, EHR audit logs can reflect the collective temporal insights of all roles (eg, physicians, nurses, laboratory staff, service assistants) involved in the treatment of a patient.

The way that the discovered 2 user-EHR interactions—visiting the order reconciliation section of the EHR system and viewing discharge instructions—contribute to predictions creates a natural question: do clinicians already know that a patient will be discharged within 24 h? It was reported in a prospective study conducted in a large academic medical center (Duke University Medical Center with 950 beds) that the recall of inpatient general medicine physicians' prediction of the next-day hospital discharge can be as low as 0.27 (morning), 0.55 (midday), and 0.67 (evening).[12] It implies that physicians might not precisely know whether their patients would discharge in the next day. Furthermore, according to our investigations at VUMC (Supplementary Appendix SF), the 2 types of user-EHR interactions are insufficient to make predictions. In other words, the presence (or absence) of these 2 interactions do not strongly associate with a next-day discharge. Rather, the discharge predictions were made in this study by learning from complicated signals generated from all care participants of a patient.

This study has several limitations that enable opportunities for future investigations. First, we acknowledge that this study is based on data from a single hospital system only. The generalizability of the findings will need to be assessed in additional healthcare systems. This is, however, a nontrivial challenge because multiple EHR systems are currently in use and EHR audit logs are not standardized among hospitals, which suggests that each hospital can leverage our framework to learn discharge prediction models relying on their own user-EHR interaction data. This will also mitigate the bias of our institution's work patterns. Workgroups from the National Research Network for Audit Log data are currently pushing EHR systems toward standardization (eg, common user-EHR interactions), but consensus takes time.[27,30] Second, this study relied on categories of user-EHR interactions as defined by Epic, which are not specifically designed for outcome prediction. It is worthwhile to investigate how different categorization criteria influence the representative power of this data source. An optimized user-EHR interaction taxonomy may enhance the performance of the prediction model. Third, this study focused on the raw sequence of user-EHR interactions and did not relate them to the corresponding clinical tasks (a higher level of concept categorizing user-EHR interactions) that clinicians collaborated to finish on the timeline. Identifying this information can potentially remove noise from the raw interaction data and provide a clearer interpretation for predictive models. Fourth, we did not explicitly represent the role of health workers in our models, which as a new dimension of EHR audit logs may enhance patient representation. Finally, in this study, we did not incorporate all data about patients in the prediction models (eg, we neglected laboratory results and medications). We believe that adding these features could further improve the prediction performance.

## FUNDING

## AUTHOR CONTRIBUTIONS

All authors contributed to study conception and design, and critical revision of the manuscript for important intellectual content. YC and XZ accessed and verified the underlying data. XZ and CY performed model development and assessment. CY and XZ drafted the manuscript. All authors contributed to the analysis and interpretation of data.

## ETHICAL APPROVAL

The institutional review board at Vanderbilt University Medical Center approved the study.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

## CONFLICT OF INTEREST STATEMENT

None declared.

## DATA AVAILABILITY

The datasets generated and/or analyzed during the current study are not publicly available due to patient private information investigated but are available from the corresponding authors on reasonable request.

## REFERENCES

1. Baru RA, Cudney EA, Guardiola IG, Warner DL, Phillips RE. Systematic review of operations research and simulation methods for bed management. In: IIE Annual Conference and Expo; 2015: 298–306.
2. Chalfin DB, Trzeciak S, Likourezos A, Baumann BM, Dellinger RP; DELAY-ED study group. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Crit Care Med* 2007; 35 (6): 1477–83.
3. Rezaei F, Yarmohammadian M, Haghshenas A, Tavakoli N. Overcrowding in emergency departments: a review of strategies to decrease future challenges. *J Res Med Sci* 2017; 22 (1): 23.
4. Al Taleb AR, Hoque M, Hasanat A, Khan MB. Application of data mining techniques to predict length of stay of stroke patients. In: 2017 International Conference on Informatics, Health & Technology (ICIHT); 2017: 1–5. doi: 10.1109/ICIHT.2017.7899004.
5. Seaton SE, Barker L, Jenkins D, Draper ES, Abrams KR, Manktelow BN. What factors predict length of stay in a neonatal unit: a systematic review. *BMJ Open* 2016; 6 (10): e010466.
6. Wiratmadja II, Salamah SY, Govindaraju R. Healthcare data mining: predicting hospital length of stay of dengue patients. *J Eng Technol Sci* 2018; 50 (1): 110.
7. Levin S, Barnes S, Toerper M, *et al.* Machine-learning-based hospital discharge predictions can support multidisciplinary rounds and decrease hospital length-of-stay. *BMJ Innov* 2021; 7 (2): 414–21.
8. Temple MW, Lehmann CU, Fabbri D. Predicting discharge dates from the neonatal intensive care unit using progress note data. *Pediatrics* 2015; 136 (2): e395–405.
9. Hisham S, Rasheed SA, Dsouza B. Application of predictive modelling to improve the discharge process in hospitals. *Healthc Inform Res* 2020; 26 (3): 166–74.
10. Gramaje A, Thabtah F, Abdelhamid N, Ray SK. Patient discharge classification using machine learning techniques. *Ann Data Sci* 2019. doi: 10.1007/s40745-019-00223-6.
11. Hintz SR, Bann CM, Ambalavanan N, Cotten CM, Das A, Higgins RD; for the Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network. Predicting time to hospital discharge for extremely preterm infants. *Pediatrics* 2010; 125 (1): e146–54.
12. Sullivan B, Ming D, Boggan JC, *et al.* An evaluation of physician predictions of discharge on a general medicine service: physician predictions of discharge. *J Hosp Med* 2015; 10 (12): 808–10.
13. Meo N, Paul E, Wilson C, Powers J, Magbual M, Miles KM. Introducing an electronic tracking tool into daily multidisciplinary discharge rounds on a medicine service: a quality improvement project to reduce length of stay. *BMJ Open Qual* 2018; 7 (3): e000174.
14. De Grood A, Blades K, Pendharkar SR. A review of discharge-prediction processes in acute care hospitals. *Healthc Policy* 2016; 12 (2): 105–15.
15. Cowan MJ, Shapiro M, Hays RD, *et al.* The effect of a multidisciplinary hospitalist/physician and advanced practice nurse collaboration on hospital costs. *J Nurs Adm* 2006; 36 (2): 79–85.
16. Resar R, Nolan K, Kaczynski D, Jensen K. Using real-time demand capacity management to improve hospitalwide patient flow. *Jt Comm J Qual Patient Saf* 2011; 37 (5): 217–27.
17. Barnes S, Hamrock E, Toerper M, Siddiqui S, Levin S. Real-time prediction of inpatient length of stay for discharge prioritization. *J Am Med Inform Assoc* 2016; 23 (e1): e2–10.
18. Gusmão Vicente F, Polito Lomar F, Mélot C, Vincent J-L. Can the experienced ICU physician predict ICU length of stay and outcome better than less experienced colleagues? *Intensive Care Med* 2004; 30 (4): 655–9.
19. Levin SR, Harley ET, Fackler JC, *et al.* Real-time forecasting of pediatric intensive care unit length of stay using computerized provider orders. *Crit Care Med* 2012; 40 (11): 3058–64.
20. Meyfroidt G, Güiza F, Cottem D, *et al.* Computerized prediction of intensive care unit discharge after cardiac surgery: development and validation of a Gaussian processes model. *BMC Med Inform Decis Mak* 2011; 11 (1): 64.
21. Awad A, Bader-El-Den M, McNicholas J. Patient length of stay and mortality prediction: a survey. *Health Serv Manage Res* 2017; 30 (2): 105–20.
22. Ward A, Mann A, Vallon J, Escobar G, Bambos N, Schuler A. Operationally-informed hospital-wide discharge prediction using machine learning. In: 2020 IEEE International Conference on E-Health Networking, Application & Services (HEALTHCOM); 2021: 1–6. doi: 10.1109/HEALTHCOM49281.2021.9399025.
23. Zanger J. *Predicting Surgical Inpatients' Discharges at Massachusetts General Hospital*. Cambridge, MA: Massachusetts Institute of Technology; 2018.
24. Oliveira S, Portela F, Santos M, Machado J, Abelha A. Hospital bed management support using regression data mining models. In: International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO); 2014: 1651–61.
25. Hachesu PR, Ahmadi M, Alizadeh S, Sadoughi F. Use of data mining techniques to determine and predict length of stay of cardiac patients. *Healthc Inform Res* 2013; 19 (2): 121–9.
26. Safavi KC, Khaniyev T, Copenhaver M, *et al.* Development and validation of a machine learning model to aid discharge processes for inpatient surgical care. *JAMA Netw Open* 2019; 2 (12): e1917221.
27. Adler-Milstein J, Adelman JS, Tai-Seale M, Patel VL, Dymek C. EHR audit logs: a new goldmine for health services research? *J Biomed Inform* 2020; 101: 103343.
28. Arndt BG, Beasley JW, Watkinson MD, *et al.* Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. *Ann Fam Med* 2017; 15 (5): 419–26.
29. Amroze A, Field TS, Fouayzi H, *et al.* Use of electronic health record access and audit logs to identify physician actions following noninterruptive alert opening: descriptive study. *JMIR Med Inform* 2019; 7 (1): e12650.
30. Sinsky CA, Rule A, Cohen G, *et al.* Metrics for assessing physician activity using electronic health record log data. *J Am Med Inform Assoc* 2020; 27 (4): 639–43.

31. Mai MV, Orenstein EW, Manning JD, Luberti AA, Dziorny AC. Attributing patients to pediatric residents using electronic health record features augmented with audit logs. *Appl Clin Inform* 2020; 11 (3): 442–51.

32. Kannampallil T, Abraham J, Lou SS, Payne PRO. Conceptual considerations for using EHR-based activity logs to measure clinician burnout and its effects. *J Am Med Inform Assoc* 2021; 28 (5): 1032–7.

33. Chen B, Alrifai W, Gao C, et al. Mining tasks and task characteristics from electronic health record audit logs with unsupervised machine learning. *J Am Med Inform Assoc* 2021. 28 (6): 1168–77. doi: 10.1093/jamia/ocaa338.

34. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc* 2013; 20 (1): 117–21.

35. Hripcsak G, Albers DJ. Correlating electronic health record concepts with healthcare process events. *J Am Med Inform Assoc* 2013; 20 (e2): e311–8.

36. Rossetti SC, Knaplund C, Albers D, et al. Healthcare process modeling to phenotype clinician behaviors for exploiting the signal gain of clinical expertise (HPM-ExpertSignals): development and evaluation of a conceptual framework. *J Am Med Inform Assoc* 2021; 28 (6): 1242–51.

37. Hribar MR, Read-Brown S, Goldstein IH, et al. Secondary use of electronic health record data for clinical workflow analysis. *J Am Med Inform Assoc* 2018; 25 (1): 40–6.

38. Jones B, Zhang X, Malin BA, Chen Y. Learning tasks of pediatric providers from Electronic Health Record audit logs. *AMIA Annu Symp Proc* 2020; 2020: 612–8.

39. Nestor JG, Fedotov A, Fasel D, et al. An electronic health record (EHR) log analysis shows limited clinician engagement with unsolicited genetic test results. *JAMIA Open* 2021; 4 (1): ooab014.

40. Wu DTY, Smart N, Ciemins EL, Lanham HJ, Lindberg C, Zheng K. Using EHR audit trail logs to analyze clinical workflow: a case study from community-based ambulatory clinics. *AMIA Annu Symp Proc* 2017; 2017: 1820–7.

41. Rule A, Chiang MF, Hribar MR. Using electronic health record audit logs to study clinical activity: a systematic review of aims, measures, and methods. *J Am Med Inform Assoc* 2020; 27 (3): 480–90.

42. Moy AJ, Schwartz JM, Elias J, et al. Time-motion examination of electronic health record utilization and clinician workflows indicate frequent task switching and documentation burden. *AMIA Annu Symp Proc* 2020; 2020: 886–95.

43. Verma G, Ivanov A, Benn F, et al. Analyses of electronic health records utilization in a large community hospital. *PLoS One* 2020; 15 (7): e0233004.

44. Chen Y, Malin B. Detection of anomalous insiders in collaborative environments via relational analysis of access logs. *CODASPY* 2011; 2011: 63–74.

45. Yan C, Li B, Vorobeychik Y, Laszka A, Fabbri D, Malin B. Get your workload in order: game theoretic prioritization of database auditing. In: 2018 IEEE 34th International Conference on Data Engineering (ICDE); 2018: 1304–7. doi: 10.1109/ICDE.2018.00136.

46. Hebbring SJ. The challenges, advantages and future of phenome-wide association studies. *Immunology* 2014; 141 (2): 157–65.

47. Overhage JM, McCallie D Jr. Physician time spent using the electronic health record during outpatient encounters: a descriptive study. *Ann Intern Med* 2020; 172 (3): 169–74.

48. Ke G, Meng Q, Finley T, et al. LGBM: a highly efficient gradient boosting decision tree. In: The 31st International Conference on Neural Information Processing Systems (NeurIPS); 2017: 3149–57.

49. Artzi NS, Shilo S, Hadar E, et al. Prediction of gestational diabetes based on nationwide electronic health records. *Nat Med* 2020; 26 (1): 71–6.

50. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 2019; 572 (7767): 116–9.

51. Choi E, Bahadori MT, Kulas JA, Schuetz A, Stewart WF, Sun J. RETAIN: an interpretable predictive model for healthcare using REverse Time AttentIoN mechanism [Internet]. *arXiv [cs.LG]* 2016. http://arxiv.org/abs/1608.05745 Accessed June 20, 2021.

52. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; 3 (1): 32–5.

53. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc* 2020; 27 (4): 621–33.

54. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks [Internet]. *arXiv [cs.LG]* 2017. http://arxiv.org/abs/1706.04599 Accessed May 12, 2021.

55. Lundberg S, Lee S-I. A unified approach to interpreting model predictions [Internet]. *arXiv [cs.AI]* 2017. http://arxiv.org/abs/1705.07874 Accessed March 26, 2021.

56. Sinha A, Stevens LA, Su F, Pageler NM, Tawfik DS. Measuring electronic health record use in the pediatric ICU using audit-logs and screen recordings. *Appl Clin Inform* 2021; 12 (4): 737–44.

57. Yan C, Zhang X, Gao C, et al. Collaboration structures in COVID-19 critical care: retrospective network analysis study. *JMIR Hum Factors* 2021; 8 (1): e25724.

58. Chen Y, Yan C, Patel MB. Network analysis subtleties in ICU structures and outcomes. *Am J Respir Crit Care Med* 2020; 202 (11): 1606–7.