Scalable and accurate deep learning for electronic health records

Alvin Rajkomar*[1,2], Eyal Oren*[1], Kai Chen[1], Andrew M. Dai[1], Nissan Hajaj[1], Peter J. Liu[1], Xiaobing Liu[1], Mimi Sun[1], Patrik Sundberg[1], Hector Yee[1], Kun Zhang [1], Yi Zhang[1], Gavin E. Duggan[1], Gerardo Flores[1], Michaela Hardt[1], Jamie Irvine[1], Quoc Le[1], Kurt Litsch[1], Jake Marcus[1], Alexander Mossin[1], Justin Tansuwan[1], De Wang[1], James Wexler[1], Jimbo Wilson[1], Dana Ludwig[2], Samuel L. Volchenboum[4], Katherine Chou[1], Michael Pearson[1], Srinivasan Madabushi[1], Nigam H. Shah[3], Atul J. Butte[2], Michael Howell[1], Claire Cui[1], Greg Corrado[1], and Jeff Dean[1]

[1]Google Inc, Mountain View, California
[2]University of California, San Francisco, San Francisco, California
[3]Stanford University, Stanford, California
[4]University of Chicago Medicine, Chicago, Illinois

January 2018

Abstract

Predictive modeling with electronic health record (EHR) data is anticipated to drive personalized medicine and improve healthcare quality. Constructing predictive statistical models typically requires extraction of curated predictor variables from normalized EHR data, a labor-intensive process that discards the vast majority of information in each patient's record. We propose a representation of patients' entire, raw EHR records based on the Fast Healthcare Interoperability Resources (FHIR) format. We demonstrate that deep learning methods using this representation are capable of accurately predicting multiple medical events from multiple centers without site-specific data harmonization. We validated our approach using de-identified EHR data from two U.S. academic medical centers with 216,221 adult patients hospitalized for at least 24 hours. In the sequential format we propose, this volume of EHR data unrolled into a total of 46,864,534,945 data points, including clinical notes. Deep learning models achieved high accuracy for tasks such as predicting in-hospital mortality (AUROC across sites 0.93-0.94), 30-day unplanned readmission (AUROC 0.75-0.76), prolonged length of stay (AUROC 0.85-0.86), and all of a patient's final diagnoses (frequency-weighted AUROC 0.90). These models outperformed state-of-the-art traditional predictive models in all cases. We also present a case-study of a neural-network attribution system, which illustrates how clinicians can gain some transparency into the predictions. We believe that this approach can be used to create accurate and scalable predictions for a variety of clinical scenarios, complete with explanations that directly highlight evidence in the patient's chart.

*These authors contributed equally

1 Introduction

The promise of digital medicine stems in part from the hope that, by digitizing health data, we might more easily leverage computer information systems to understand and improve care. In fact, routinely collected patient healthcare data is now approaching the genomic scale in volume and complexity [94]. Unfortunately, most of this information is not yet used in the sorts of predictive statistical models clinicians might use to improve care delivery. It is widely suspected that such efforts, if successful, could provide major benefits not only for patient safety and quality but also in reducing health care costs [7, 56, 47, 74].

In spite of the richness and potential of available data, scaling the development of predictive models is difficult because, for traditional predictive modeling techniques, each outcome to be predicted requires the creation of a custom dataset with specific variables [34]. It is widely held that 80% of the effort in an analytic model is preprocessing, merging, customizing, and cleaning data sets, [79, 63] rather than analyzing them for insights. This profoundly limits the scalability of predictive models.

Another challenge is that the number of potential predictor variables in the electronic health record (EHR) may easily number in the thousands, particularly if free-text notes from doctors, nurses, and other providers are included. Traditional modeling approaches have dealt with this complexity simply by choosing a very limited number of commonly-collected variables to consider [34]. This is problematic because the resulting models may produce imprecise predictions: false-positive predictions can overwhelm physicians, nurses, and other providers with false alarms and subsequent alert fatigue [27], which the Joint Commission identified as a national patient safety priority in 2014 [18]. False-negative predictions can miss significant numbers of clinically important events, leading to poor clinical outcomes [51]. Incorporating the *entire* EHR, including clinicians' free-text notes, offers some hope of overcoming these shortcomings but is hopelessly unwieldy for most predictive modeling techniques.

Recent developments in deep learning and artificial neural networks may allow us to address many of these challenges and unlock the information in the EHR. Deep learning emerged as the preferred machine learning approach in machine perception problems ranging from computer vision to speech recognition, but has more recently proven useful in natural language processing, sequence prediction, and mixed modality data settings([59, 32, 39, 102]). These systems are known for their ability to handle large volumes of relatively messy data, including errors in labels and large numbers of input variables. A key advantage is that investigators do not generally need to specify which potential predictor variables to consider and in what combinations; instead neural networks to learn representations of the key factors and interactions from the data itself.

We hypothesize that these techniques will translate well to healthcare. Specifically, that deep learning approaches could incorporate the entire electronic health record, including free-text notes, to produce predictions for a wide range of clinical problems and outcomes that outperformed state-of-the-art traditional predictive models. Our central insight is that rather than explicitly harmonizing EHR data, mapping it into a highly curated set of structured predictors variables and then feeding those variables into a statistical model, we can instead learn to simultaneously harmonize inputs and predict medical events through direct feature learning [9].

Related Work

Using computer systems to learn from a "highly organized and recorded database" of clinical data has a long history [100]. Despite the rich data now digitized in EHRs [3], a recent systematic review of the medical literature [34] found that predictive models built with EHR data use a median of only 27 variables, rely on traditional generalized linear models, and are built using data at a single center. In clinical practice, simpler models are most commonly deployed, such as the CURB-65 [68, 60], which is a 5-factor model, or single-parameter warning scores [20, 45].

A major challenge in using more of the data available for each patient has been the lack of standards and semantic interoperability of health data from multiple sites [91]. A unique set of variables is typically selected for each new prediction task, and usually a labor-intensive [78, 64] process is required to extract and normalize data from different sites [70].

Significant prior research has focused on the scalability issue through time-consuming standardization of data in traditional relational databases, like the Observational Medical Outcomes Partnership (OMOP) standard defined by the Observational Health Data Sciences and Informatics (OHDSI) consortium [73]. Such a standard allows for consistent development of predictive models across sites, but accommodates only a part of the original data.

Recently, a flexible data structure called Fast Healthcare Interoperability Resources (FHIR) [66] was developed to represent clinical data in a consistent, hierarchical, and extensible container format, regardless of the health system, which simplifies data interchange between sites. However, the format does not ensure semantic consistency, increasing the need for additional techniques to deal with unharmonized data.

The use of deep learning on electronic health record data burgeoned after adoption of electronic health records [3] and development of deep learning methods [59]. In a well-known work, investigators used auto-encoders to predict a specific set of diagnoses [69]. Subsequent work extended this approach by modeling the temporal sequence of events that occurred in a patient's record, which may enhance accuracy in scenarios that depend on the order of events, with convolutional and recurrent neural networks [61, 2, 17, 81, 16, 92]. In general, prior work has focused on a subset of features available in the EHR, rather than on all data available in an electronic health record, which includes clinical free-text notes as well as large amounts of structured and semi-structured data. Because of the availability of Medical Information Mart for Intensive Care (MIMIC) data [48], many prior studies also have focused on ICU patients from a single center [92, 41]; other single-center studies have also focused on ICU patients [61]. Each ICU patient has significantly more data available than each general hospital patient, although non-ICU admissions outnumber ICU admissions by about 6-fold in the US [24, 30]. Recently, investigators have explored how interpretation mechanisms for deep learning models could be applied to clinical predictions [92]. Given rapid developments in this field, we point readers to a recent, comprehensive review [88].

Our contribution is two-fold. First, we report a generic data processing pipeline that can take raw EHR data as input, and produce FHIR outputs without manual feature harmonization. This makes it relatively easy to deploy our system to a new hospital. Second, based on data from two academic hospitals with a general patient population (not restricted to ICU), we demonstrate the effectiveness of using deep learning models in a wide variety of predictive problems and settings (e.g. multiple prediction timing). Ours is a comprehensive study of deep learning in a variety of prediction problems based on multiple general hospital data. We do note, however, that similar deep learning techniques have been applied to EHR data in prior research as described above.

2 Methods

Datasets

We included EHR data from the University of California, San Francisco (UCSF) from 2012-2016, and the University of Chicago Medicine (UCM) from 2009-2016. We refer to each health system as Hospital A and Hospital B. All electronic health records were de-identified, except that dates of service were maintained in the UCM dataset. Both datasets contained patient demographics, provider orders, diagnoses, procedures, medications, laboratory values, vital signs, and flowsheet data, which represents all other structured data elements (e.g. nursing flowsheets), from all inpatient and outpatient encounters. The UCM dataset (but not UCSF) additionally contained de-identified, free-text medical notes. Each dataset was kept in an encrypted, access-controlled, and audited sandbox.

Ethics review and institutional review boards approved the study with waiver of informed consent or exemption at each institution.

Data representation and processing

We developed a single data structure that could be used for any prediction, rather than requiring custom, hand-created datasets for every new prediction. This approach represents the entire EHR in temporal order: data are organized by patient and by time. To represent events in a patient's timeline, we adopted the FHIR standard (Fast Healthcare Interoperability Resources) [43]. FHIR defines the high-level representation of healthcare data into resources, but leaves values in each individual site's idiosyncratic codings [67]. Each event is derived from a FHIR resource and may contain multiple attributes; for example a medication-order resource could contain the trade name, generic name, ingredients, and others. Data in each attribute was split into discrete values which we refer to as a tokens. For notes, the text was split into a sequence of tokens, one for each word. Numeric values were normalized, as detailed in the appendix. The entire sequence of time-ordered tokens, from the beginning of a patient's record until the point of prediction, formed the patient's personalized input to the model. This process is illustrated in Figure 1.

Outcomes

We were interested in understanding whether deep learning could produce valid predictions across wide range of clinical problems and outcomes. We therefore selected outcomes from divergent domains, including an important clinical outcome (death), a standard measure of quality of care (readmissions), a measure of resource utilization (length of stay), and a measure of understanding of a patient's problems (diagnoses).

Inpatient mortality We predicted impending inpatient death, defined as a discharge disposition of "expired" [93, 52, 97, 103].

30-day unplanned readmission We predicted unplanned 30 day readmission, defined as an admission within 30 days after discharge from an "index" hospitalization. A hospitalization was considered a "readmission" if admission date was within thirty days after discharge of an eligible index hospitalization. A readmission could only be counted once. There is no standard definition of "unplanned" [28] so we used a modified form of the Centers for Medicare and Medicaid Services (CMS) definition [1] which we detail in the appendix. Billing diagnoses and procedures from the

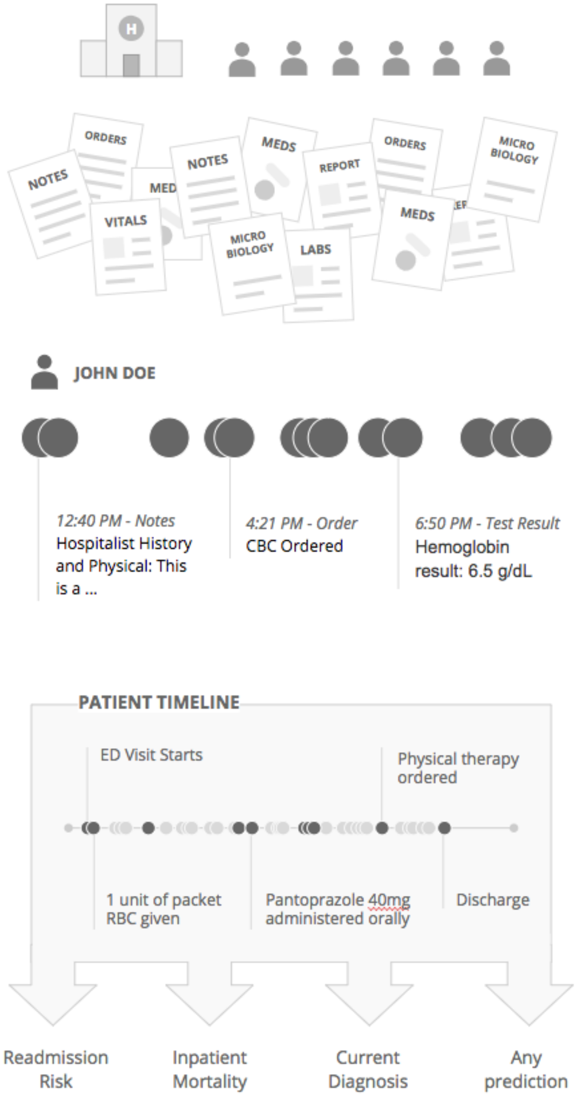

1
Health systems collect and store electronic health records in various formats in databases.
ORDERS
NOTES
MED
VITALS
NOTES
MEDS
MICRO BIOLOGY
LABS
REPORT
ORDERS
MEDS
MICRO BIOLOGY
2
All available data for each patient is converted to events recorded in containers based on the Fast Healthcare Interoperability Resource (FHIR) specification.
JOHN DOE
12:40 PM - Notes
Hospitalist History and Physical: This is a ...
4:21 PM - Order
CBC Ordered
6:50 PM - Test Result
Hemoglobin result: 6.5 g/dL
3
The FHIR resources are placed in temporal order, depicting all events recorded in the EHR (i.e. timeline). The deep learning model uses this full history to make each prediction.
PATIENT TIMELINE
ED Visit Starts
Physical therapy ordered
1 unit of packet RBC given
Pantoprazole 40mg administered orally
Discharge
Readmission Risk
Inpatient Mortality
Current Diagnosis
Any prediction

Figure 1: Data from each health system an appropriate FHIR (Fast Healthcare Interoperability Resources) resource and placed in temporal order. This conversion did not harmonize or standardize the data from each health-system other than map them to the appropriate resource. The deep learning model could use all data available prior to the point when the prediction was made. Therefore each prediction, regardless of the task, used the same data.

index hospitalization were not used for the prediction because they are typically generated after discharge. We included only readmissions to the same institution.

Long length of stay We predicted a length-of-stay at least 7 days, which was approximately the 75th percentile hospital stays for most services across the datasets. The length-of-stay was defined as the time between hospital admission and discharge.

Diagnoses We predicted the entire set of primary and secondary ICD-9 billing diagnoses (i.e. from a universe of 14,025 codes).

Prediction Timing

This was a retrospective study. To predict inpatient mortality, we stepped forward through each patient's time course, and made predictions every twelve hours starting 24 hours before admission until 24 hours after admission. Since many clinical prediction models, such as APACHE, are rendered 24 hours after admission, our primary outcome prediction for in-patient mortality was at that time-point. Unplanned readmission and the set of diagnosis codes were predicted at admission, 24 hours after admission, and at discharge. The primary endpoints for those predictions were at discharge, when most readmission prediction scores are computed and when all information necessary to assign billing diagnoses is available [50]. Long-length of stay was predicted at admission and 24 hours after admission. For every prediction we used all information available in the EHR up to the time at which the prediction was made.

Study Cohort We included all consecutive admissions for patients 18 years or older. We only included hospitalizations of 24 hours or longer to ensure that predictions at various timepoints had identical cohorts.

To simulate the accuracy of a real-time prediction system, we included patients typically removed in studies of readmission, such as those discharged against medical advice, since these exclusion criteria would not be known when making predictions earlier in the hospitalization.

For predicting the ICD-9 diagnoses, we excluded encounters without any ICD-9 diagnosis (2-12% of encounters). These were generally encounters after October, 2015 when hospitals switched to ICD-10. We included such hospitalizations, however, for all other predictions.

Algorithm Development and Analysis We used the same modeling algorithm on both hospitals' datasets, but treated each hospital as a separate dataset and report results separately.

Patient records vary significantly in length and density of data-points (e.g. vital sign measurements in an intensive care unit vs outpatient clinic), so we formulated three deep learning neural-network model architectures that take advantage of such data in different ways: one based on recurrent neural networks (LSTM) [44], one on an attention-based time-aware neural network model (TANN), and one on a neural network with boosted time-based decision stumps. Details of these architectures are explained in the appendix. Each model was trained on each hospital's data separately for each prediction and time-point. To optimize model accuracy, our final model is an ensemble of predictions from the three underlying model architectures [82].

Comparison to previously published algorithms We implemented models based on previously published algorithms to establish baseline performance on each dataset. For mortality, we used a logistic model with variables inspired by NEWS [89] score but added additional variables to make it more accurate, including the most recent systolic blood pressure, heart rate, respiratory rate, temperature, and we used 24 common lab tests, like the white-blood cell count, lactate, and creatinine. We call this the augmented Early-Warning Score, or aEWS, score. For readmission, we used a logistic model with variables used by the HOSPITAL [25] score, including the most recent sodium and hemoglobin level, hospital service, occurrence of CPT codes, number of prior hospitalizations, and length of the current hospitalization. We refer to this as the mHOSPITAL score. For long length of stay, we used a logistic model with variables similar to those used by Liu [62]: the age, gender, Hierarchical Condition Categories, admission source, hospital service, and the same 24 common lab tests used in the aEWS score. We refer to this as the modified Liu (mLiu) score. Details for the baseline models are in the appendix. We are not aware of any commonly used baseline model for all diagnosis codes so we compare against known literature.

Explanation of predictions A common criticism of neural networks is that they offer little insight into the factors that influence the prediction [13]. Therefore, we used attribution mechanisms to highlight, for each patient, the data elements that influenced their predictions [5].

The LSTM and TANN models were trained with TensorFlow and the boosting model was implemented with C++ code. Statistical analyses and baseline models were done in Scikit-learn Python [75].

Technical details of the model architecture, training, variables, baseline models, and attribution methods are provided in the appendix.

Model Evaluation and Statistical Analysis Patients were randomly split into development (80%), validation (10%) and test (10%) sets. Model accuracy is reported on the test set, and 1000 bootstrapped samples were used to calculate 95% confidence intervals. To prevent overfitting, the test set remained unused (and hidden) until final evaluation.

We assessed model discrimination by calculating area under the receiver operating characteristic curve (AUROC) and model calibration using comparisons of predicted and empirical probability curves.(Pencina and D'Agostino 2015) We did not use the Hosmer-Lemeshow test as it may be misleadingly significant with large sample sizes [55]. To quantify the potential clinical impact of an alert with 80% sensitivity, we report the work-up to detection ratio, also known as the number needed to evaluate [83]. For prediction of the a patient's full set of diagnosis codes, which can range from 1 to 228 codes per hospitalization, we evaluated the accuracy for each class using macro-weighted-AUROC [85] and micro-weighted F1 score [86] to compare with the literature. The F1 score is the harmonic mean of positive-predictive-value and sensitivity; we used a single threshold picked on the validation set for all classes. We did not create confidence intervals for this task given the computational complexity of the number of possible diagnoses.

3 Results

We included a total of 216,221 hospitalizations involving 114,003 unique patients. The percent of hospitalizations with in-hospital deaths was 2.3% (4,930/216,221), unplanned 30-day readmissions was 12.9% (27,918/216,221), and long length of stay was (23.9%). Patients had a range of 1 to 228 discharge diagnoses. Demographics and utilization characteristics are summarized in Table 1. At the time of admission, an average admission had 137,882 tokens, which increased markedly throughout the patient's stay to 216,744 at discharge (Figure 2). For predictions made at discharge, the information considered across both datasets included 46,864,534,945 tokens of EHR data.

Mortality For predicting inpatient mortality, the AUROC at 24 hours after admission was 0.95 (95% CI 0.94-0.96) for Hospital A and 0.93 (95% CI 0.92-0.94) for Hospital B. This was significantly more accurate than the traditional predictive model, aEWS which was a 28-factor logistic regression model (AUROC 0.85 [95% CI 0.81-0.89] for Hospital A and 0.86 [95% CI 0.83-0.88] for Hospital B).

Table 1: Characteristics of Hospitalizations in Training and Test Sets

	Training Data (n=194,470)		Test Data (n=21,751)	
	Hospital A (n=85,522)	Hospital B (n=108,948)	Hospital A (n=9,624)	Hospital B (n=12,127)
Demographics				
Age, median (IQR) y	56 (29)	57 (29)	55 (29)	57 (30)
Female sex, No. (%)	46 848(54.8%)	62 004(56.9%)	5364(55.7%)	6935(57.2%)
Disease Cohort, No (%)				
Medical	46 579(54.5%)	55 087(50.6%)	5263(54.7%)	6112(50.4%)
Cardiovascular	4616 (5.4%)	6903 (6.3%)	528 (5.5%)	749 (6.2%)
Cardiopulmonary	3498 (4.1%)	9028 (8.3%)	388 (4.0%)	1102 (9.1%)
Neurology	6247 (7.3%)	6653 (6.1%)	697 (7.2%)	736 (6.1%)
Cancer	14 544(17.0%)	19 328(17.7%)	1617(16.8%)	2087(17.2%)
Psychiatry	788 (0.9%)	339 (0.3%)	64 (0.7%)	35 (0.3%)
Obstetrics & newborn	8997(10.5%)	10 462 (9.6%)	1036(10.8%)	1184 (9.8%)
Other	253 (0.3%)	1148 (1.1%)	31 (0.3%)	122 (1.0%)
Previous Hospitalizations				
0 hospitalizations	54 954(64.3%)	56 197(51.6%)	6123(63.6%)	6194(51.1%)
$\geq$ 1 and $<$ 2 hospitalizations	14 522(17.0%)	19 807(18.2%)	1620(16.8%)	2175(17.9%)
$\geq$ 2 and $<$ 6 hospitalizations	12 591(14.7%)	24 009(22.0%)	1412(14.7%)	2638(21.8%)
$\geq$ 6 hospitalizations	3455 (4.0%)	8935 (8.2%)	469 (4.9%)	1120 (9.2%)
Discharge Location				
Home	70 040(81.9%)	91 273(83.8%)	7938(82.5%)	10 109(83.4%)
Skilled Nursing Facility	6601 (7.7%)	5594 (5.1%)	720 (7.5%)	622 (5.1%)
Rehabilitation	2666 (3.1%)	5136 (4.7%)	312 (3.2%)	649 (5.4%)
Another Healthcare Facility	2189 (2.6%)	2052 (1.9%)	243 (2.5%)	220 (1.8%)
Expired	1816 (2.1%)	2679 (2.5%)	170 (1.8%)	265 (2.2%)
Other	2210 (2.6%)	2214 (2.0%)	241 (2.5%)	262 (2.2%)
Primary Outcomes				
In-hospital deaths, No. (%)	1816 (2.1%)	2679 (2.5%)	170 (1.8%)	265 (2.2%)
30-day readmissions No. (%)	9136(10.7%)	15 932(14.6%)	1013(10.5%)	1837(15.1%)
Hospital stays at least 7 days, No. (%)	20 411(23.9%)	26 109(24.0%)	2145(22.3%)	2931(24.2%)

Table 2: Prediction Accuracy of Each Task Made at Different Time Points

	Hospital A	Hospital B
Inpatient Mortality, AUROC[1](95% CI)		
24 hours before admission	0.87 (0.85-0.89)	0.81 (0.79-0.83)
At admission	0.90 (0.88-0.92)	0.90 (0.86-0.91)
24 hours after admission	**0.95**(0.94-0.96)	**0.93**(0.92-0.94)
Baseline (aEWS[2]) at 24 hours after admission	0.85 (0.81-0.89)	0.86 (0.83-0.88)
30-day Readmission, AUROC (95% CI)		
At admission	0.73 (0.71-0.74)	0.72 (0.71-0.73)
24 hours after admission	0.74 (0.72-0.75)	0.73 (0.72-0.74)
At discharge	**0.75**(0.75-0.78)	**0.76**(0.75-0.77)
Baseline (mHOSPITAL[3]) at discharge	0.70 (0.68-0.72)	0.68 (0.67-0.69)
Length of Stay at least 7 days AUROC (95% CI)		
At admission	0.81 (0.80-0.82)	0.80 (0.80-0.81)
24 hours after admission	**0.86**(0.86-0.87)	**0.85**(0.85-0.86)
Baseline (mLiu[4]) at 24 hours after admission	0.76 (0.75-0.77)	0.74 (0.73-0.75)
Discharge Diagnoses, (weighted AUROC)		
At admission	0.87	0.86
24 hours after admission	0.89	0.88
At discharge	**0.90**	**0.90**

[1] Area under the receiver operator curve
[2] augmented early warning score
[3] modified HOSPITAL score
[4] modified Liu score

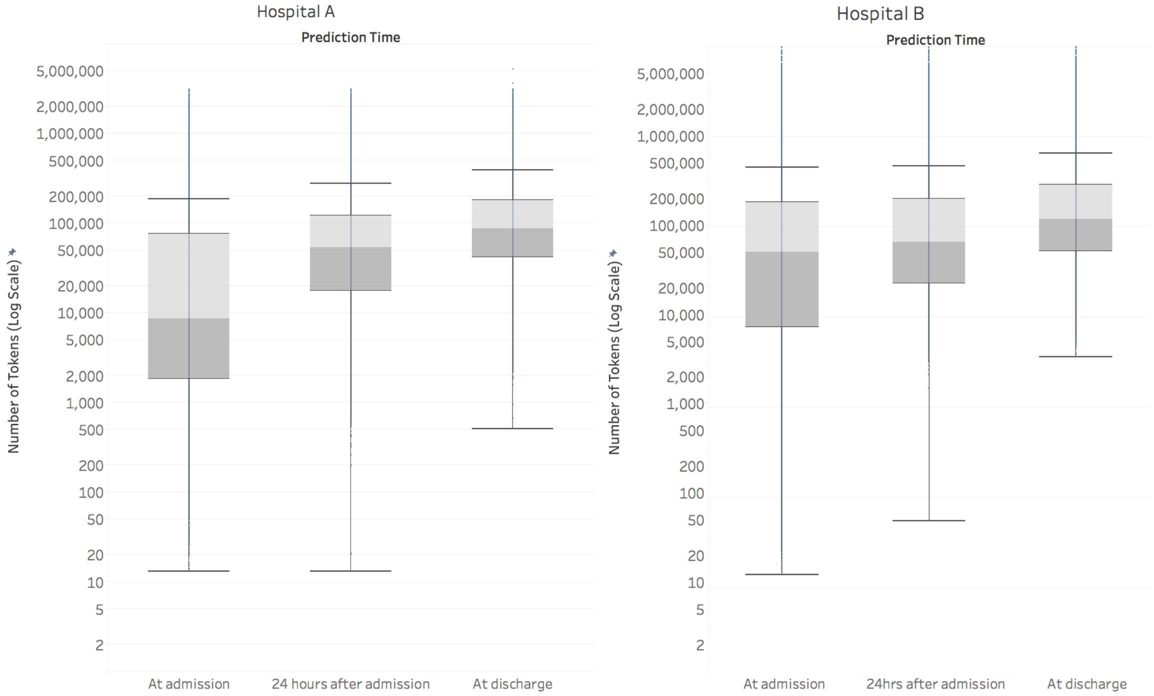


Figure 2: This boxplot displays the amount of data (on a log scale) in the EHR, along with its temporal variation across the course of an admission. We define a token as a single data element in the electronic health record, like a medication name, at a specific point in time. Each token is considered as a potential predictor by the deep learning model. The line within the boxplot represents the median, the box represents the interquartile range (IQR), and the whiskers are 1.5 times the IQR. The number of tokens increased steadily from admission to discharge. At discharge, the median number of tokens for Hospital A was 86,477 and for Hospital B was 122,961.

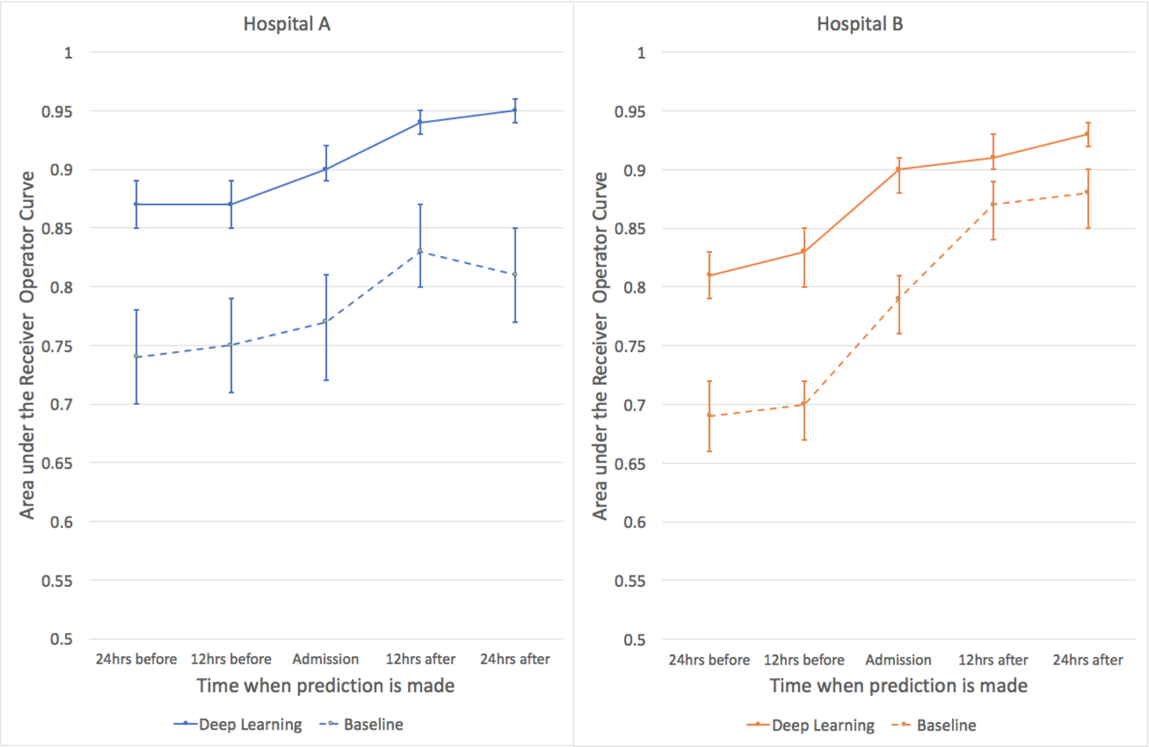


Figure 3: The area under the receiver operator curves are shown for predictions of inpatient mortality made by deep learning and baseline models at twelve hour increments before and after hospital admission. For inpatient mortality, the deep learning model achieves higher discrimination at every prediction time compared to the baseline for both the University of California, San Francisco (UCSF) and University of Chicago Medicine (UCM) cohorts. Both models improve in the first 24 hours, but the deep learning model achieves a similar level of accuracy approximately 24 hours earlier for UCM and even 48 hours earlier for UCSF. The error bars represent the bootstrapped 95% confidence interval.

If a clinical team had to investigate patients predicted to be at high risk of dying, the rate of false alerts at each point in time was roughly halved by our model: at 24 hours, the work-up to detection ratio of our model compared to the aEWS was 7.4 vs 14.3 (Hospital A) and 8.0 vs 15.4 (Hospital B). The deep learning model predicted events 24-48 hours earlier than the traditional predictive model (Figure 3).

Readmissions For predicting unexpected readmissions within 30-days, the AUROCs at discharge were 0.77 (95% CI 0.75-0.78) for Hospital A and 0.76 (95% CI 0.75-0.77) for Hospital B. These were significantly higher than the traditional predictive model (mHOSPITAL) model at discharge, which were 0.70 (95% CI 0.68-0.72) for Hospital A and 0.68 (95%CI 0.67-0.69) for Hospital B.

Long Length of Stay For predicting long length-of-stay, the AUROCs at 24 hours after admission were 0.86 (95% CI 0.86-0.87) for Hospital A and 0.85 (95% CI 0.84-0.86) for Hospital B. These were significantly higher than the traditional predictive model In (mLiu) at 24 hours, which were 0.76 (95% CI 0.75- 0.77) for Hospital A and 0.74 (95% CI 0.73-0.75) for Hospital B.

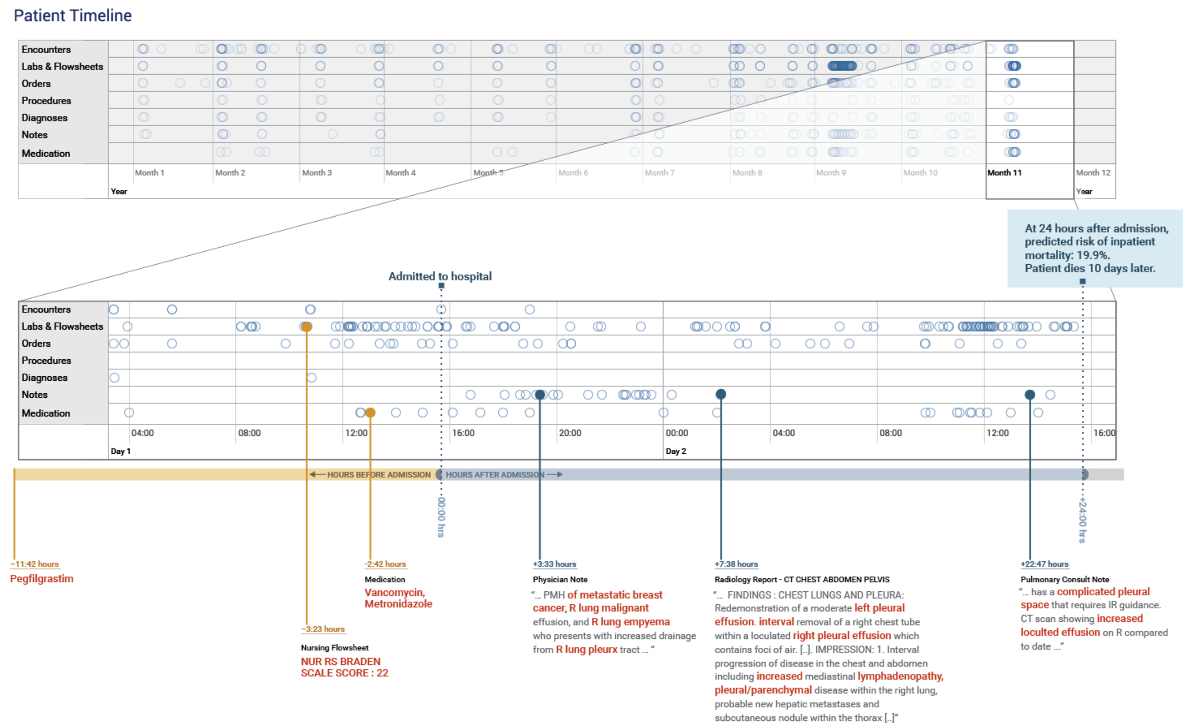
Calibration curves for the three tasks are shown in the appendix.

Inferring Discharge Diagnoses The deep learning algorithm predicted patients' discharge diagnoses at three time points: at admission, after 24 hours of hospitalization, and at the time of discharge (but before the discharge diagnoses were coded). For classifying all diagnosis codes, the weighted AUROCs at admission were 0.87 for Hospital A and 0.86 for Hospital B. Accuracy increased somewhat during the hospitalization, to 0.88-0.89 at 24 hours and 0.90 for both hospitals at discharge. For classifying ICD-9 code predictions as correct, we required full-length code agreement. For example, 250.4 ("Diabetes with renal manifestations") would be considered different from 250.42 ("Diabetes with renal manifestations, type II or unspecified type, uncontrolled"). We also calculated the micro-F1 score at discharge which were 0.41 (Hospital A) and 0.40 (Hospital B).

Case study of Model Interpretation In Figure 4, we illustrate an example of attribution methods on a specific prediction of inpatient-mortality made at 24 hours after admission. For this patient, the deep learning model predicted the risk of death of 19.9% and the baseline model predicted 9.3%, and the patient ultimately died 10 days after admission. This patient's record had 175,639 data points (tokens) which were considered by the model. The timeline in Figure 4 highlights the elements to which the model attends, with a close-up view of the first 24 hours of the most recent hospitalization. From all the data, the models picked the elements that are highlighted in Figure 4: evidence of malignant pleural effusions and empyema from notes, antibiotics administered, and nursing documentation of a high risk of pressure ulcers (e.g. Braden index).(Bergstrom et al. 1987) The model also placed high weights on concepts such as "pleurx," the trade-name for a small chest tube. The bolded sections are exactly what the model identified as discriminatory factors, not a manual selection. In contrast, the top predictors for the baseline model (not shown in Figure 4) were the values of the albumin, blood-urea-nitrogen, pulse, and white blood cell count. Note that for demonstration purposes, this example was generated from TANNs trained on separate modalities (e.g. flowsheets and notes), which is a common visualization technique to handle redundant features in the data (e.g. medication orders are also referenced in notes).

4 Discussion

A deep learning approach that incorporated the entire electronic health record, including free-text notes, produced predictions for a wide range of clinical problems and outcomes that outperformed



At 24 hours after admission, predicted risk of inpatient mortality: 19.9%. Patient dies 10 days later.

−11:42 hours
Pegfilgrastim

-2:42 hours
Medication
Vancomycin, Metronidazole

−3:23 hours
Nursing Flowsheet
NUR RS BRADEN SCALE SCORE : 22

+3:33 hours
Physician Note
"... PMH of metastatic breast cancer, R lung malignant effusion, and R lung empyema who presents with increased drainage from R lung pleurx tract ... "

+7:38 hours
Radiology Report - CT CHEST ABDOMEN PELVIS
"... FINDINGS : CHEST LUNGS AND PLEURA: Redemonstration of a moderate left pleural effusion. interval removal of a right chest tube within a loculated right pleural effusion which contains foci of air. [..]. IMPRESSION: 1. Interval progression of disease in the chest and abdomen including increased mediastinal lymphadenopathy, pleural/parenchymal disease within the right lung, probable new hepatic metastases and subcutaneous nodule within the thorax [..]"

+22:47 hours
Pulmonary Consult Note
"... has a complicated pleural space that requires IR guidance. CT scan showing increased loculted effusion on R compared to date ..."

Figure 4: The patient record shows a woman with metastatic breast cancer with malignant pleural effusions and empyema. The patient timeline at the top of the figure contains circles for every time-step for which at least a single token exists for the patient, and the horizontal lines show the data-type. There is a close-up view of the most recent data-points immediately preceding a prediction made 24 hours after admission. We trained models for each data-type and highlighted in red the tokens which the models attended to – the non-highlighted text was not attended to but is shown for context. The models pick up features in the medications, nursing flowsheets, and clinical notes to make the prediction.

state-of-the-art traditional predictive models. Because we were interested in understanding whether deep learning could scale to produce valid predictions across divergent healthcare domains, we used a single data structure to make predictions for an important clinical outcome (death), a standard measure of quality of care (readmissions), a measure of resource utilization (length of stay), and a measure of understanding of a patient's problems (diagnoses).

This method represents an important advance in the scalability of predictive models in clinical care for several reasons. First, our study's approach uses a single data-representation of the entire EHR as a sequence of events, allowing this system to be used for any prediction that would be clinically or operationally useful with minimal additional data preparation. Traditional predictive models require substantial work to prepare a hand-crafted, tailored dataset with specific variables, selected by experts and assembled by analysts for each new prediction [34]. This data preparation and cleaning typically consumes up to 80% of the effort of any predictive analytics project [79, 63], limiting the scalability of predictive models in healthcare.

Second, using the entirety of a patient's chart for every prediction does more than promote scalability, it exposes more data with which to make an accurate prediction. For predictions made at discharge, our deep learning models considered more than 46 billion pieces of EHR data and achieved more accurate predictions, earlier in the hospital stay, than did traditional models.

To the best of our knowledge, our models outperform existing EHR literature for predicting mortality (0.92-0.94 vs 0.91), [93] unexpected readmission (0.75-0.76 vs 0.69)[71] and increased length of stay (0.85-0.86 vs 0.77) [62]. Direct comparisons to other studies are difficult [98] because of different underlying study designs [52, 29, 97, 103, 77, 21, 80, 89, 53, 14, 84, 31, 104, 87, 96] incomplete definitions of cohorts and outcomes[19, 15], restrictions on disease-specific cohorts [19, 15, 95, 10, 101, 23, 37] or use of data unavailable in real-time [23, 42, 33]. Therefore, we implemented baselines based on the HOSPITAL score [25], NEWS [89] score, and Liu's model [62] on our data, and demonstrate strictly better performance. We are not aware of a study that predicts as many ICD codes as this study, but our micro-F1 score exceeds that shown on the smaller MIMIC-III dataset when predicting fewer diagnoses (0.40 vs 0.28) [76]. The clinical impact of this improvement is suggested, for example, by the improvement of number needed to evaluate for inpatient mortality: the deep learning model would fire half the number of alerts of a traditional predictive model, resulting in many fewer false positives.

However, the novelty of the approach does not lie simply in incremental model performance improvements. Rather, this predictive performance was achieved without hand-selection of variables deemed important by an expert, similar to other applications of deep learning to EHR data. Instead, our model had access to tens of thousands of predictors for each patient, including free-text notes, and learned what was important for a particular prediction.

Our study also has important limitations. First, it is a retrospective study, with all of the usual limitations. Second, although it is widely believed that accurate predictions can be used to improve care [7], this is not a foregone conclusion and prospective trials are needed to demonstrate this [57, 38]. Third, a necessary implication of personalized predictions is that they leverage many small data points specific to a particular EHR rather than a handful of common variables. Future research is needed to determine how models trained at one site can be best applied to another site, [40] which would be especially useful for sites with limited historical data for training. As a first step, we demonstrated that similar model architectures and training methods yielded comparable models for two geographically distinct health systems, but further research is needed on this point. Finally, our methods are computationally intensive and at present require specialized expertise to implement. However, the availability and accessibility of machine learning is rapidly expanding both in healthcare and in other fields.

Perhaps the most challenging prediction in our study is that of predicting a patient's full suite of discharge diagnoses. The prediction is difficult for several reasons. First, a patient may have between 1 and 228 diagnoses, and the number is not known at the time of prediction. Second, each diagnosis may be selected from approximately 14,025 ICD-9 diagnoses codes, which makes the total number of possible combinations exponentially large. Finally, many ICD-9 codes are clinically similar but numerically distinct (for example, 011.30 "Tuberculosis of bronchus, unspecified" vs. 011.31 "Tuberculosis of bronchus, bacteriological or histological examination not done"). This has the effect of introducing random error into the prediction. The micro-F1 score, which is a metric used when a prediction has more than a single outcome (e.g. multiple diagnoses), for our model is higher than that reported in the literature in an ICU data-set with fewer diagnoses [76]. This is a proof-of-concept that demonstrates that the diagnosis could be inferred from routine EHR data, which could aid with triggering of decision support [8] or clinical trial recruitment.

The use of free text for prediction also allows a new level of explainability of predictions. Clinicians have historically distrusted neural network models because of their opaqueness. We demonstrate how our method can visualize what data the model "looked at" for each individual patient, which can be used by a clinician to determine if a prediction was based on credible facts, and potentially help decide actions. In our case study, the model identified elements of the patient's history and radiology findings to render its prediction, which are critical data-points that a clinician would also use [72]. This approach may address concerns that such "black box" methods are untrustworthy. However, there are other possible techniques for interpretability of deep learning models [4, 92], and further research is needed regarding both the cognitive impact of this approach and its clinical utility.

5 Conclusions

Accurate predictive models can be built directly from EHR data for a variety of important clinical problems with explanations highlighting evidence in the patient's chart.

6 Acknowledgements

For data acquisition and validation, we thank the following: Julie Johnson, Sharon Markman, Thomas Sutton, Brian Furner, Timothy Holper, Sharat Israni, Jeff Love, Doris Wong, Michael Halaas, Juan Banda. For statistical consultation, we thank Farzan Rohani. For modeling infrastructure, we thank Daniel Hurt and Zhifeng Chen. For help with visualizing Figures 1 and 3, we thank Kaye Mao and Mahima Pushkarna. We thank Zachary Lipton for helpful discussion on an earlier draft of this preprint.

A Data Representation

Data from each electronic health record was imported into a new schema based on the open-source Fast Healthcare Interoperability Resources (FHIR) resource standards. We populated relevant data into elements from the following resources: Patient, Encounter, Medication, Observation (e.g. vital signs and nursing documentation), Composition (e.g. notes), Conditions (i.e. diagnoses), MedicationAdministration, MedicationOrder, ProcedureRequest, and Procedure. We imported the data directly from the health system, meaning we did not harmonize elements to a standard terminology or ontology. If a health system included multiple terminologies, like a site-specific coding scheme and an RxNorm code (a common medication coding scheme), we imported both. The only exceptions were for diagnoses/procedures, which we mapped to ICD9/10 and CCS categories if the health system did not already include them (e.g. for CPT codes), and for elements that were used to define the primary outcomes, as described in the main manuscript.

In the electronic health record datasets, there was a category of data referred to as ”flowsheets,” which correspond to many structured data elements in clinical care, like vital signs and nursing documentation. Depending on workflows, data may be collected at the bedside, like a temperature reading and then entered in the EHR later. This documentation provides (at least) two timestamps – when the data was technically collected (recorded time) and when it was entered (entry time). We specifically used the entry time in the EHR because especially during emergent situations, the recorded times are estimated. We found that using the recorded-times significantly improved prediction accuracy, but refrained from using them as the data is not actually available in the EHR at that point-in-time.

For each categorical variable in each set of resources, we created a d-dimensional floating-point embedding vector, E, with d picked as a hyperparameter. For clinical notes, we created sequences of embeddings for words that appeared at least m times, with m as a hyperparameter. All embeddings are randomly initialized. For numeric variables, we also normalized the values. We used hyperparameter tuning to select the size of the buckets. We also did tuning to select the best way to represent specific values as combinations of embeddings representing the nearest neighbors (e.g. linear combinations of the nearest intervals). These embeddings were randomly initialized and updated over the course of training the model.

Embeddings representing all data prior to a prediction point were placed in chronological order $E_i, i = 1, ..., n$ where n is the number of elements in the sequence Each embedding vector was concatenated with a time-delta value or embedding, Δ_i representing the difference in time from the data element occurring to the time the prediction was made.

B Description of Inclusion Criteria and Outcomes

Inclusion Criteria

Inpatient encounters were defined as followed: 1. Encounter was confirmed as complete or non-cancelled 2. Encounter had a start and end time 3. Encounter class was defined as inpatient as defined in dataset 4. Administrative encounters were excluded (e.g. no primary diagnosis was documented); these encounters accounted for less than 1 percent of hospitalizations in the data received.

The following services were included in the Medical-Surgical Cohort: General Medicine, Cardiology, Neurology, Critical Care Medicine, Hematology, Oncology, Hepatobiliary Medicine, Medical Specialties, General Surgery, Colorectal Surgery, Otolaryngology, Gynecology, Gynecology-Oncology, Neurosurgery, Oral-Maxillofacial surgery, Orthopedics, Plastic Surgery, Thoracic Surgery, Transplant Surgery, Urology, and Vascular Surgery.

Cohort Definitions

We made the following modifications of the CMS cohort definitions[1] to ensure that every primary diagnosis was listed in a cohort.

We added CCS code 150 (alcoholic liver disease) to the Medicine Cohort.

We created the following new Cohorts: Obstetric Cohort containing CCS codes 176-196. Rehabilitation Cohort containing CCS code 254 Injury and Poisoning Cohort containing CCS codes 260 and 2601-2621.

Determining Unplanned Readmission

We implemented the logic used by CMS to define planned readmissions[1]. The logic evaluates whether admissions were for reasons that are defined to be planned (e.g. bone marrow transplants and chemotherapy), and it distinguishes between surgical procedures that were accompanied by an acute condition (e.g. acute cholecystitis) or non-acute condition, which were defined to be unplanned or planned, respectively.

In the 2016 version of the CMS rules, some criteria were defined by a mix of CCS and ICD-9 procedure and diagnosis codes. Given that some hospitalizations only had ICD-10 diagnoses and procedure codes, we mapped the ICD-9 CMS codes to ICD-10 and then applied the rules. We used the mapping tables provided by the National Bureau of Economic Research[22].

Fewer than 1 percent of hospitalizations did not have a primary diagnosis marked in the raw data. Based on a review of a random sample of these hospitalizations, these encounters lacked clinical data about events in the hospitalization and were therefore not included but they did have administrative data that indicated that these were unplanned admissions. After confirming with the respective partner sites, we treated these cases as ineligible to be index discharges given missing data but were considered unplanned admissions. They were excluded from the mortality and diagnosis prediction tasks.

C Model Variants

Weighted Recurrent neural network model (RNN)

In the RNN model, sparse features of each category (such as medication or procedures) were embedded into the same d-dimensional embedding. d for each category was chosen based on the number of possible features for that category. The embeddings from different categories are concatenated and for the same category and same time, they are averaged according to an automatically learned weighting.

The sequence of embeddings were further reduced down to a shorter sequence. Typically, the shorter sequences were split into time-steps of 12 hours where the embeddings for all features within a category in the same day were combined using weighted averaging. The weighted averaging is done by associating each feature with a non-negative weight that is trained jointly with the model. These weights are also used for prediction attribution. The log of the average time-delta at each time-step is also embedded into a small floating-point vector (which is also randomly initialized) and concatenated to the input embedding at each time-step.

This reduced sequence of embeddings were then fed to an n-layer Recurrent Neural Network (RNN), specifically a Long Short-Term Memory network (LSTM).3 An RNN consists of a sequence of directed nodes. Embeddings are fed to the RNN one at a time and for each time-step, each node computes its activation as a nonlinear function of the input embedding. Each subsequent node receives as input the previous node's activation and the embedding for that time-step. The LSTM extends the RNN by adding 3 gates, an input gate, output gate, forget gate to determine what information to pass on to the next node relative to the previous node's activation and the current time-step's embedding. Each node in the LSTM computes an hidden state vector and cell state vector.

The LSTM is defined by the following set of equations where W and U corresponds to weight matrices, b to biases and the subscript and variable f, i, o represent the forget, input and output gates. h_t represents the hidden output at time t, x_t represents the input at time t and c_t represents the cell state at time t. σ_g represents the sigmoid function and σ_c the hyperbolic tangent.

$$\begin{aligned} f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\ i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\ o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\ c_t &= f_t \cdot c_{t-1} + i_t \cdot \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\ h_t &= o_t \cdot \sigma_c(c_t) \end{aligned} \tag{1}$$

The hidden state of the final time-step of the LSTM was fed into an output layer, and the model was optimized to minimize the log-loss (either a logistic regression or softmax loss depending on the task). We applied a number of regularization techniques to the model, namely embedding dropout, feature dropout, LSTM input dropout and variational RNN dropout.4 We also used a small level of L2 weight decay, which adds a penalty for large weights into the loss. We trained with a batch size of 128 and clipped the norm of the gradients to 50. Finally, we optimized everything jointly with Adagrad6. We trained using the TensorFlow framework on the Tesla P100 GPU. The regularization hyperparameters and learning rate were found via a Gaussian-process based hyperparameter search on each dataset's validation performance.

Feedforward Model with Time-Aware Attention

To the sequence of embedding, $E_i, i = 1, \ldots, n$ we added an additional prior embedding to the sequence E_0 with the associated $\Delta_0 = 0$. For every embedding $E_i, i = 0, \ldots, n$ we created an attribution logit α_i using the process described below. Those logits were converted to weights β_i using softmax,

$$\beta_i = \frac{e^{\alpha_i}}{\sum_{j=0}^{n} e^{\alpha_j}} \tag{2}$$

We then took the d dimensional vector of the weighted sum, $E = \sum_j \beta_j E_j$, along with the scalars $\log(n+1)$ and $\log(\sum_j \beta_j^2)$, and entered them into a feedforward neural network whose attributes (e.g. number and dimensions of the layers) were determined by hyperparameter tuning.

For the attribution logits, we used a bank of k functions, $A_1(\Delta), \ldots, A_k(\Delta)$, where each A_j had one of the following forms (typically not all forms in the same model):

- $A(\Delta) = 1$ (constant);
- $A(\Delta) = \Delta$;
- $A(\Delta) = \log(\Delta + 1\text{day})$;
- $A(\Delta) =$ Piece-wise linear function with predetermined inflection points (based on exponential backoff) and learned slopes.

We defined a k dimensional projection of the embedding by learning a $k \times d$ dimensional matrix P, and for every $i = 0, 1, \ldots, n$ multiplying it with E_i to get the k scalars $p_{1,i}, \ldots, p_{k,i}$. We then defined the attribution logits to be

$$\alpha_i = \sum_{j=1}^{k} p_{j,i} A_j(\Delta_i) \tag{3}$$

The embedding dimension, d, ranged from 16 to 512. The number of layers of the feedforward network ranged from 0-3 with the width of the networks from 10 to 512.

Boosted, embedded time-series model

For each feature tuple of the token name, value, and time-delta, we algorithmically created (described below) a set of binary decision rules that partitioned examples into two classes.

There were ten types of decision rules.

- The first was whether a variable, X, existed at any-point in a patient's timeline.
- The second was whether a variable X existed more than C times in a patient's timeline. C was randomly picked from the range of integer values possible for each variable in the dataset.
- The third introduced the time sequence nature of the variable: was variable X greater or lower than threshold V at any time $t <$ T (i.e. $x > V$ and $t < T$; or $x \leq V$ and $t < T$). Again, V and T were picked from the space of possible values in the dataset.

  The fourth was a modification of the third rule, but rather than a simple binary cutoff, it was a weighted sum of of the number of times that rule ($x < V$) was satisfied, with the weights determined by a Hawkes process response with a time decay factor of T. A binary rule was created by examining if this weighted sum was greater than the activation A, that is $A_{instance} > A_{template}$, where $A_{template}$ is selected from a random user. Again, we use random selection of a particular template instance to select V and T. Then A is computed from the instance by

$$A = \sum_i I\{x_i > V\} e^{\frac{-t_i}{T}} \tag{4}$$

- The fifth, six, and seventh rules were created by determining if the minimum, maximum, and average of variable X was greater than V in time $t < T$.
- The eighth and ninth type of rule captured changes in lab values over time (e.g. the decrease in blood pressure over time). In particular, the eighth predicate checked the velocity, that is if there is a change in a variable divided by a time window that is greater or lower than a threshold V. The ninth predicate checked if the difference in values within time T is greater or lower than a threshold V.
- The tenth type of rule consists of conjunctions of previous predicates (e.g. does X_1 exist and does the count of X_2 exceed C_2),We call these decision list predicates as, to preserve interpretability, they only encompass the true branches of a decision tree. The conjunctions are mined by picking the best predicate in a random selection of predicates, then, conditioned on

the best predicate, the a second one that also maximizes the weighted information gain with respect to the label.

The actual instances of each rules, including the selection of variables, value thresholds and time-thresholds were generated by first picking a random patient, random variable X, and a random time T in the patient's timeline. V is the corresponding value of X at time T and C is the counts of times X occurs in the patient's timeline.

Every binary rule, which we refer to as a binary predicate, was assigned a scalar weight, and the weighted sum was passed through a softmax layer to create a prediction. To train, we first created a bias binary predicate which was true for all examples and its weight was assigned as the log-odds ratio of the positive label class across the dataset.

Next, we used rounds of boosting to select predicates. In each round, we picked 25,000 random predicates from random patients in a batch of 500 patients. Importance-weighted information gain with respect to the label was calculated for each and the top 50 predicates were picked. Additionally, for each of those top 50 predicates, 50 more secondary predicates were selected using the same information gain criteria, conditional on the primary predicate holding true. The best predicate and second corresponding predicates were then joined together to create 50 more conjunction predicates for a total of 100 predicates per round. Weights of these predicates were fitted using logistic regression with L_1 regularization. We then applied the model to all examples in the training dataset to create prediction probabilities Q. Each example was then given an importance weight of $|Label - Q|$.

In the next round, we selected 25,000 new random predicates by sampling examples according to the importance weight. The top 50 by information gain (and 50 more secondary ones) were added to a new logistic model which included the predicates from the previously determined predicates. The weights of all predicates were re-calculated (i.e. not just the new predicates), which is known as totally corrective boosting.

We used 100 rounds, so in total 100,000 predicates were selected from a pool of 5,000,000 which were in turn randomly selected from a potential pool of $num\_patient * num\_features * num\_discrete\_values * num\_time\_steps$ potentially possible predicates. The L_1 regularization was then used, which could further cull away the 100,000 selected predicates to a smaller set.

The final binary predicates were then embedded into a 1024 dimensional vector space and then fed to a feed-forward network of depth 2 and 1024 hidden units per layer with ELU non-linearity. For regularization, Gaussian noise of mean 0 and standard deviation 0.8 was added to the input of the feed forward network. We also used multiplicative Bernoulli noise of p=0.01 (also known as dropout) at the input and output (just before the applying the sigmoid function to the logits) of the feed forward layer. At test time, no Gaussian or Bernoulli noise was used. We optimized everything with Adam. The union of predicates optimized for different tasks (e.g. readmission or different diagnosis codes), were all used together in the final model. These final binary predicates have been mined from different tasks (e.g. for the readmission task, many diagnosis code models might contribute auxiliary binary predicates that they have mined as features for the feedforward network).

D Methods for All Techniques

Attribution Mechanisms

To explain predictions we implement attribution mechanisms. Inspired by recent results in natural language processing [6], the feed-forward models implement an attention mechanism identifying the locations in a sequence of variables which may have played a significant role in affecting the prediction. Notably, the same variable could be harnessed differently given when it occurred in relationship to other events for a given patient timeline. The RNN models implement a form of weighting that also learns which variables are important for prediction relative to other variables. We use both of these methods to perform attribution.

Illustrating the data that the models attended to is difficult because of the complexity of the data, including thousands of time-steps with tens- to hundreds-of-thousands of tokens, representing a large percentage of all the data that is viewable in a patient's actual EHR record. Moreover, given the correlation of the data (the heart-rate at time x is related to the rate at $x + 1$) and redundancy (the medication order of "norepinpherine" is redundant with the nurse's documentation of the rate to which it is titrated), the models could choose to attend to equivalent data arbitrarily. For visualization purposes only, we re-trained feed-forward models on a single task, mortality, with models using only a single data-type (e.g. notes, medications, observation data) to preclude the models using redundant data among different feature types. These models differ in predictive performance than that of the full models reported in Table 2.

In Figure 4 of the main manuscript, we render the timeline, populating a circle for every time-step where at least a single token exists for that patient. We have shown snippets of select time-steps and highlight the tokens in which the model using that data-type chose to attend it. For tokens with significant attribution scores, we have "smeared" attribution to directly neighboring tokens for visualization purposes. For patient privacy reasons, we have obscured information about the dates and times of all tokens, although the relative time has been retained.

Automated Hyperparameter Tuning

There are many design choices to training neural networks that are beyond the scope of this manuscript but are well described elsewhere [36]. The hyper-parameters, which are settings that affect the performance of all above neural networks were tuned automatically using Google Vizier [35] with a total of >201,000 GPU hours.

Ensembling

For a given prediction task, we could use a variety of algorithms to make a prediction. For example, we could use a sequence model, feed-forward model, and a boosting model, and their predictions would be different on the same example. Ensembling combines the multiple predictions to make a final prediction; this is similar to tallying votes for an election result. We combined the predictions (probabilities) from the three models of the ensemble by averaging.

E Baseline Models

To understand the performance of our models, we first created baseline models for each prediction task using traditional modeling techniques. We used recent literature reviews to select commonly used variables for each task [105, 90, 65]. These hand-engineered features are used only in the baseline models; our actual models do not need such feature engineering.

We fitted the model on the training set separately for each planet and report results when applying this model to the test set of each planet.

Mortality Baseline Model - aEWS

Most existing models use a small set of lab measurements, vital signs and mental status. Following this approach, for the EHR datasets, we created a model that used the most recent systolic blood pressure, heart-rate, respiratory rate and temperature in fahrenheit (any temperature that was listed below 90 degree fahrenheit was converted from Celsius to Fahrenheit). Because urine output and mental status was not coded consistently between sites, we instead used the most recent white blood cell count, hemoglobin [49], sodium [11], creatinine [12, 58], troponin [99, 54, 46] lactate oxygen saturation, oxygen source, glucose, calcium, potassium, chloride, blood urea nitrogen (BUN), carbon dioxide, hematocrit, platelet, magnesium, phosphorus, albumin, aspartate transaminase (AST), Alkaline Phosphatase, Total Bilirubin, International Normalized Ratio, and Absolute Neutrophil Count (ANC). All values were log transformed and standardized to have a mean of zero and standard deviation of 1 based on values for each planet on the development set. We also added the hospital service and age.

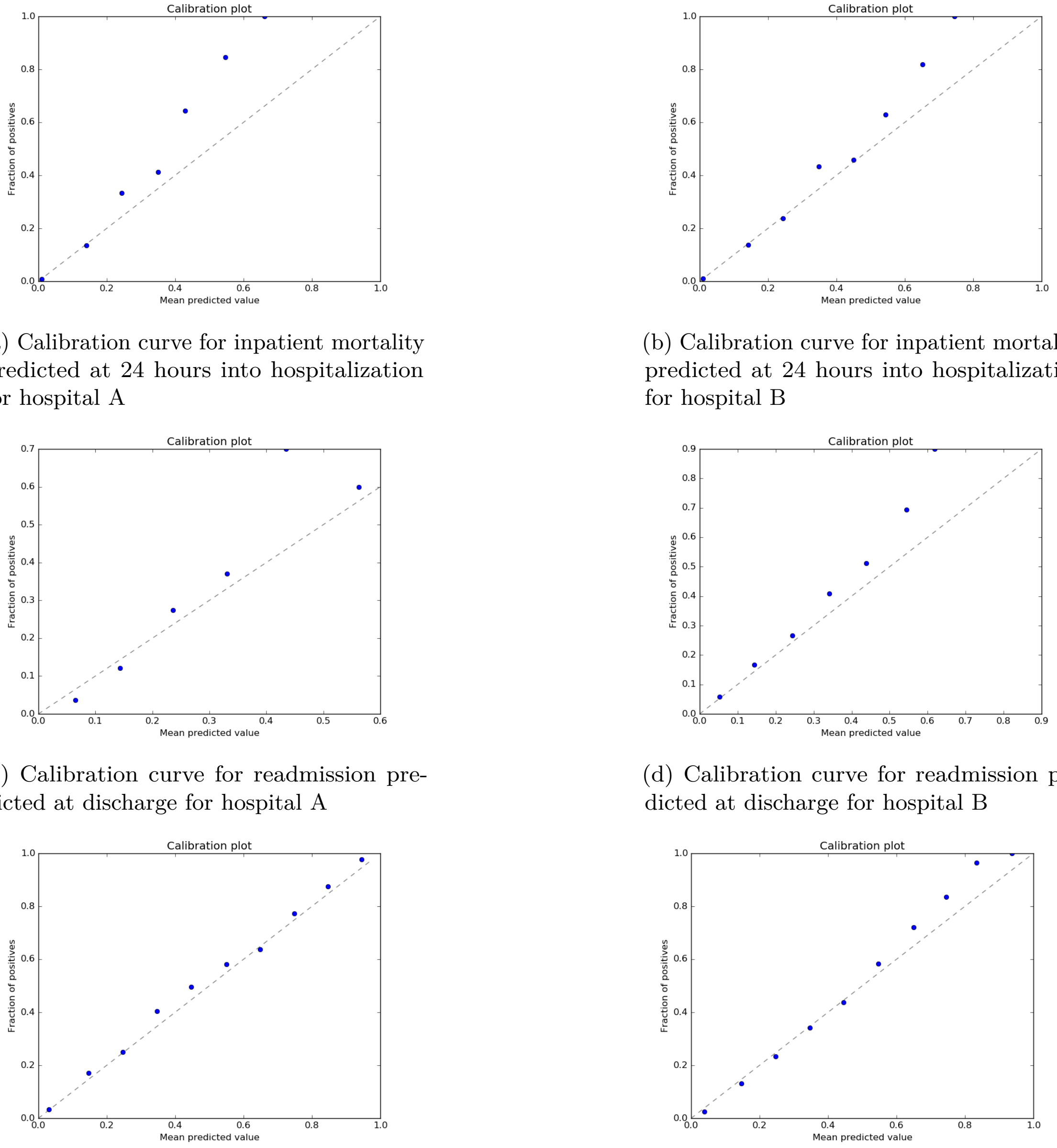
Readmission Baseline Model - modified HOSPITAL score

We created a modified HOSPITAL score [26] that included the most recent value of sodium and hemoglobin log transformed and standardized (to mean 0 and standard deviation of 1) based on values for each planet on the development set, binary indicators for hospital service, a binary indicator for the occurrence of any CPT codes during the hospitalization, a binary indicator for the hospitalization lasting at least 5 days, prior hospital admissions in the past year discretized to 0,1, 2-5 and >5, and admission source.

Length of Stay Baseline Model - modified Liu

We created a baseline model similar to those created using electronic health record data for general hospital populations18,19. In particular, we created a lasso logistic model with the following variables: age, gender, HCC (Hierarchical Condition Categories) codes in the past year (counts for each one), admission source, hospital service, and the lab predictors from the mortality baseline model.

We chose to create a baseline model similar to those created using electronic health record data [62]. We created a lasso logistic model with the following variables: age, gender, prior HCC codes in the timeline (counts for each one), the principal diagnosis coded as a CCS, hospital service, and the most recent lab value of each possible lab.



(a) Calibration curve for inpatient mortality predicted at 24 hours into hospitalization for hospital A

(b) Calibration curve for inpatient mortality predicted at 24 hours into hospitalization for hospital B

(c) Calibration curve for readmission predicted at discharge for hospital A

(d) Calibration curve for readmission predicted at discharge for hospital B

(e) Calibration curve for long length of stay predicted at 24 hours into hospitalization for hospital A

(f) Calibration curve for long length of stay predicted at 24 hours into hospitalization for hospital B

References

[1] "2016 Measure updates and specifications report: hospital-wide all-cause unplanned readmission — version 5.0". In: *Yale–New Haven Health Services Corporation/Center for Outcomes Research & Evaluation* (May 2016).

[2] M Aczon et al. "Dynamic Mortality Risk Predictions in Pediatric Critical Care Using Recurrent Neural Networks". In: (Jan. 2017). arXiv: `1701.06675 [stat.ML]`.

[3] Julia Adler-Milstein et al. "Electronic Health Record Adoption In US Hospitals: Progress Continues, But Challenges Persist". en. In: *Health Aff.* 34.12 (Dec. 2015), pp. 2174–2180.

[4] Anand Avati et al. "Improving Palliative Care with Deep Learning". In: (Nov. 2017). arXiv: `1711.06402 [cs.CY]`.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: (Sept. 2014). arXiv: `1409.0473 [cs.CL]`.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: (Jan. 2014). arXiv: `1409.0473 [cs.CL]`.

[7] D W Bates et al. "Big data in health care: using analytics to identify and manage high-risk and high-cost patients". In: *Health Aff.* 33.7 (2014), pp. 1123–1131.

[8] David W Bates et al. "Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality". en. In: *J. Am. Med. Inform. Assoc.* 10.6 (Nov. 2003), pp. 523–530.

[9] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation Learning: A Review and New Perspectives". In: (June 2012). arXiv: `1206.5538 [cs.LG]`.

[10] Vasiliki Betihavas et al. "An Absolute Risk Prediction Model to Determine Unplanned Cardiovascular Readmissions for Adults with Chronic Heart Failure". en. In: *Heart Lung Circ.* 24.11 (Nov. 2015), pp. 1068–1073.

[11] Scott W Biggins et al. "Serum sodium predicts mortality in patients listed for liver transplantation". en. In: *Hepatology* 41.1 (Jan. 2005), pp. 32–39.

[12] Ion D Bucaloiu et al. "Increased risk of death and de novo chronic kidney disease following reversible acute kidney injury". en. In: *Kidney Int.* 81.5 (Mar. 2012), pp. 477–485.

[13] Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini. "Unintended Consequences of Machine Learning in Medicine". In: *JAMA* (July 2017).

[14] Xiongcai Cai et al. "Real-time prediction of mortality, readmission, and length of stay using electronic health record data". en. In: *J. Am. Med. Inform. Assoc.* 23.3 (May 2016), pp. 553–561.

[15] Rich Caruana et al. "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD '15. New York, NY, USA: ACM, 2015, pp. 1721–1730.

[16] Zhengping Che et al. "Recurrent Neural Networks for Multivariate Time Series with Missing Values". In: (June 2016). arXiv: `1606.01865 [cs.LG]`.

[17] Edward Choi et al. "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks". In: *Proceedings of the 1st Machine Learning for Healthcare Conference.* jmlr.org, 2016, pp. 301–318.

[18] Vineet Chopra and Laurence F McMahon Jr. "Redesigning hospital alarms for patient safety: alarmed and potentially dangerous". en. In: *JAMA* 311.12 (Mar. 2014), pp. 1199–1200.

[19] Shahid A Choudhry et al. "A public-private partnership develops and externally validates a 30-day hospital readmission risk prediction model". en. In: *Online J. Public Health Inform.* 5.2 (July 2013), p. 219.

[20] Matthew M Churpek and Dana P Edelson. "Moving Beyond Single-Parameter Early Warning Scores for Rapid Response System Activation". en. In: *Crit. Care Med.* 44.12 (Dec. 2016), pp. 2283–2285.

[21] Matthew M Churpek et al. "Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards". en. In: *Crit. Care Med.* 44.2 (Feb. 2016), pp. 368–374.

[22] *CMS' ICD-9-CM to and from ICD-10-CM and ICD-10-PCS Crosswalk or General Equivalence Mappings.* `http://www.nber.org/data/icd9-icd-10-cm-and-pcs-crosswalk-general-equivalence-mapping.html`. Accessed: 2017-7-21.

[23] Eric A Coleman et al. "Posthospital care transitions: patterns, complications, and risk identification". en. In: *Health Serv. Res.* 39.5 (Oct. 2004), pp. 1449–1465.

[24] *Critical Care Statistics.* `http://www.sccm.org/Communications/Pages/CriticalCareStats.aspx`. Accessed: 2018-1-25.

[25] Jacques Donzé et al. "Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model". en. In: *JAMA Intern. Med.* 173.8 (Apr. 2013), pp. 632–638.

[26] Jacques Donzé et al. "Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model". en. In: *JAMA Intern. Med.* 173.8 (22 4 2013), pp. 632–638.

[27] Barbara J Drew et al. "Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients". en. In: *PLoS One* 9.10 (Oct. 2014), e110274.

[28] Gabriel J Escobar et al. "Nonelective Rehospitalizations and Postdischarge Mortality: Predictive Models Suitable for Use in Real Time". en. In: *Med. Care* 53.11 (Nov. 2015), pp. 916–923.

[29] Gabriel J Escobar et al. "Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases". en. In: *Med. Care* 46.3 (Mar. 2008), pp. 232–239.

[30] *Fast Facts on U.S. Hospitals, 2018.* `https://www.aha.org/statistics/fast-facts-us-hospitals`. Accessed: 2018-1-25.

[31] G Duncan Finlay, Michael J Rothman, and Robert A Smith. "Measuring the modified early warning score and the Rothman index: advantages of utilizing the electronic medical record in an early warning system". en. In: *J. Hosp. Med.* 9.2 (Feb. 2014), pp. 116–119.

[32] Andrea Frome et al. “DeViSE: A Deep Visual-Semantic Embedding Model”. In: *Advances in Neural Information Processing Systems 26*. Ed. by C J C Burges et al. Curran Associates, Inc., 2013, pp. 2121–2129.

[33] Joseph Futoma, Jonathan Morris, and Joseph Lucas. “A comparison of models for predicting early hospital readmissions”. en. In: *J. Biomed. Inform.* 56 (Aug. 2015), pp. 229–238.

[34] Benjamin A Goldstein et al. “Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review”. en. In: *J. Am. Med. Inform. Assoc.* 24.1 (Jan. 2017), pp. 198–208.

[35] Daniel Golovin et al. “Google Vizier: A Service for Black-Box Optimization”. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.

[36] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[37] Evan M Graboyes et al. “Risk Factors for Unplanned Hospital Readmission in Otolaryngology Patients”. en. In: *Otolaryngol. Head Neck Surg.* (Aug. 2013).

[38] Kevin Grumbach, Catherine R Lucey, and S Claiborne Johnston. “Transforming From Centers of Learning to Learning Health Systems: The Challenge for Academic Health Centers”. In: *JAMA* 311.11 (Mar. 2014), pp. 1109–1110.

[39] Varun Gulshan et al. “Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs”. In: *JAMA* 316.22 (Dec. 2016), pp. 2402–2410.

[40] John D Halamka and Micky Tripathi. “The HITECH Era in Retrospect”. en. In: *N. Engl. J. Med.* 377.10 (Sept. 2017), pp. 907–909.

[41] Hrayr Harutyunyan et al. “Multitask Learning and Benchmarking with Clinical Time Series Data”. In: (Mar. 2017). arXiv: `1703.07771 [stat.ML]`.

[42] Danning He et al. “Mining high-dimensional administrative claims data to predict early hospital readmissions”. en. In: *J. Am. Med. Inform. Assoc.* 21.2 (Mar. 2014), pp. 272–279.

[43] *Health Level 7*. `http://hl7.org/fhir/`. Accessed: 2017-8-3. Apr. 2017.

[44] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780.

[45] Michael D Howell et al. “Sustained effectiveness of a primary-team-based rapid response system”. en. In: *Crit. Care Med.* 40.9 (Sept. 2012), pp. 2562–2568.

[46] P James et al. “Relation between troponin T concentration and mortality in patients presenting with an acute stroke: observational study”. en. In: *BMJ* 320.7248 (Mar. 2000), pp. 1502–1504.

[47] J Larry Jameson and Dan L Longo. “Precision medicine–personalized, problematic, and promising”. en. In: *N. Engl. J. Med.* 372.23 (June 2015), pp. 2229–2234.

[48] Alistair E W Johnson et al. “MIMIC-III, a freely accessible critical care database”. en. In: *Sci Data* 3 (May 2016), p. 160035.

[49] Paul R Kalra et al. “Hemoglobin and Change in Hemoglobin Status Predict Mortality, Cardiovascular Events, and Bleeding in Stable Coronary Artery Disease”. en. In: *Am. J. Med.* (19 1 2017).

[50] Devan Kansagara et al. “Risk prediction models for hospital readmission: a systematic review”. en. In: *JAMA* 306.15 (Oct. 2011), pp. 1688–1698.

[51] Kirsi-Maija Kaukonen et al. “Systemic inflammatory response syndrome criteria in defining severe sepsis”. en. In: *N. Engl. J. Med.* 372.17 (Apr. 2015), pp. 1629–1638.

[52] John Kellett and Arnold Kim. “Validation of an abbreviated Vitalpac$^{\text{TM}}$ Early Warning Score (ViEWS) in 75,419 consecutive admissions to a Canadian regional hospital”. en. In: *Resuscitation* 83.3 (Mar. 2012), pp. 297–302.

[53] Hargobind S Khurana et al. “Real-Time Automated Sampling of Electronic Medical Records Predicts Hospital Mortality”. en. In: *Am. J. Med.* 129.7 (July 2016), 688–698.e2.

[54] Lauren J Kim et al. “Cardiac troponin I predicts short-term mortality in vascular surgery patients”. en. In: *Circulation* 106.18 (29 10 2002), pp. 2366–2371.

[55] Andrew A Kramer and Jack E Zimmerman. “Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited”. en. In: *Crit. Care Med.* 35.9 (Sept. 2007), pp. 2052–2056.

[56] Harlan M Krumholz. “Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system”. en. In: *Health Aff.* 33.7 (July 2014), pp. 1163–1170.

[57] Harlan M Krumholz, Sharon F Terry, and Joanne Waldstreicher. “Data Acquisition, Curation, and Use for a Continuously Learning Health System”. en. In: *JAMA* 316.16 (Oct. 2016), pp. 1669–1670.

[58] Jean-Philippe Lafrance and Donald R Miller. “Acute kidney injury associates with increased long-term mortality”. en. In: *J. Am. Soc. Nephrol.* 21.2 (Feb. 2010), pp. 345–352.

[59] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. en. In: *Nature* 521.7553 (May 2015), pp. 436–444.

[60] W S Lim et al. “British Thoracic Society community acquired pneumonia guideline and the NICE pneumonia guideline: how they fit together”. en. In: *BMJ Open Respir Res* 2.1 (May 2015), e000091.

[61] Zachary C Lipton et al. “Learning to Diagnose with LSTM Recurrent Neural Networks”. In: (Nov. 2015). arXiv: `1511.03677 [cs.LG]`.

[62] Vincent Liu et al. “Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables”. en. In: *Med. Care* 48.8 (Aug. 2010), pp. 739–744.

[63] Steve Lohr. “For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights”. In: *The New York Times* (Aug. 2014).

[64] Steve Lohr. “For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights”. In: *The New York Times* (Aug. 2014).

[65] Mingshan Lu et al. “Systematic review of risk adjustment models of hospital length of stay (LOS)”. en. In: *Med. Care* 53.4 (Apr. 2015), pp. 355–365.

[66] Joshua C Mandel et al. “SMART on FHIR: a standards-based, interoperable apps platform for electronic health records”. en. In: *J. Am. Med. Inform. Assoc.* 23.5 (Sept. 2016), pp. 899–908.

[67] Joshua C Mandel et al. “SMART on FHIR: a standards-based, interoperable apps platform for electronic health records”. en. In: *J. Am. Med. Inform. Assoc.* 23.5 (Sept. 2016), pp. 899–908.

[68] Lionel A Mandell et al. “Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults”. en. In: *Clin. Infect. Dis.* 44 Suppl 2 (Mar. 2007), S27–72.

[69] Riccardo Miotto et al. “Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records”. en. In: *Sci. Rep.* 6 (May 2016), p. 26094.

[70] Katherine M Newton et al. “Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network”. en. In: *J. Am. Med. Inform. Assoc.* 20.e1 (June 2013), e147–54.

[71] Oanh Kieu Nguyen et al. “Predicting all-cause readmissions using electronic health record data from the entire hospitalization: Model development and comparison”. en. In: *J. Hosp. Med.* 11.7 (July 2016), pp. 473–480.

[72] Ziad Obermeyer and Ezekiel J Emanuel. “Predicting the Future — Big Data, Machine Learning, and Clinical Medicine”. In: *N. Engl. J. Med.* 375.13 (2016), pp. 1216–1219.

[73] *OMOP Common Data Model.* `https://www.ohdsi.org/data-standardization/the-common-data-model/`. Accessed: 2018-1-23.

[74] Ravi B Parikh, J Sanford Schwartz, and Amol S Navathe. “Beyond Genes and Molecules - A Precision Delivery Initiative for Precision Medicine”. en. In: *N. Engl. J. Med.* 376.17 (Apr. 2017), pp. 1609–1612.

[75] Fabian Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *J. Mach. Learn. Res.* 12.Oct (2011), pp. 2825–2830.

[76] Adler Perotte et al. “Diagnosis code assignment: models and evaluation metrics”. en. In: *J. Am. Med. Inform. Assoc.* 21.2 (Mar. 2014), pp. 231–237.

[77] Michael Pine et al. “Modifying ICD-9-CM coding of secondary diagnoses to improve risk-adjustment of inpatient mortality rates”. en. In: *Med. Decis. Making* 29.1 (Jan. 2009), pp. 69–81.

[78] G Press - Forbes, March, and 2016. “Cleaning big data: Most time-consuming, least enjoyable data science task, survey says”. In: (2016).

[79] Gil Press. *Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says.* `https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/`. Accessed: 2017-10-22. Mar. 2016.

[80] David R Prytherch et al. “ViEWS–Towards a national early warning score for detecting adult inpatient deterioration”. en. In: *Resuscitation* 81.8 (Aug. 2010), pp. 932–937.

[81] Narges Razavian, Jake Marcus, and David Sontag. “Multi-task Prediction of Disease Onsets from Longitudinal Lab Tests”. In: (Aug. 2016). arXiv: `1608.00647 [cs.LG]`.

[82] Lior Rokach. “Ensemble-based Classifiers”. In: *Artif. Intell. Rev.* 33.1-2 (Feb. 2010), pp. 1–39.

[83] Santiago Romero-Brufau et al. “Why the C-statistic is not informative to evaluate early warning scores and what metrics to use”. In: *Crit. Care* 19.1 (2015), p. 285.

[84] Michael J Rothman, Steven I Rothman, and Joseph Beals 4th. “Development and validation of a continuous measure of patient condition using the Electronic Medical Record”. en. In: *J. Biomed. Inform.* 46.5 (Oct. 2013), pp. 837–848.

[85] *SciKit Learn Documentation on Area Under the Curve scores.* `http://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html`. Accessed: 2017-8-3.

[86] *SciKit Learn Documentation on F1 Score.* `http://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html`. Accessed: 2017-8-3.

[87] Issac Shams, Saeede Ajorlou, and Kai Yang. "A predictive analytics approach to reducing avoidable hospital readmission". In: (Feb. 2014). arXiv: `1402.5991 [stat.AP]`.

[88] Benjamin Shickel et al. "Deep EHR: A Survey of Recent Advances on Deep Learning Techniques for Electronic Health Record (EHR) Analysis". In: (June 2017). arXiv: `1706.03446 [cs.LG]`.

[89] Gary B Smith et al. "The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death". en. In: *Resuscitation* 84.4 (Apr. 2013), pp. 465–470.

[90] M E Beth Smith et al. "Early Warning System Scores for Clinical Deterioration in Hospitalized Patients: A Systematic Review". In: *Ann. Am. Thorac. Soc.* 11.9 (2014), pp. 1454–1465.

[91] Hong Sun et al. "Semantic processing of EHR data for clinical research". en. In: *J. Biomed. Inform.* 58 (Dec. 2015), pp. 247–259.

[92] Harini Suresh et al. "Clinical Intervention Prediction and Understanding using Deep Networks". In: (May 2017). arXiv: `1705.08498 [cs.LG]`.

[93] Ying P Tabak et al. "Using electronic health record data to develop inpatient mortality predictive model: Acute Laboratory Risk of Mortality Score (ALaRMS)". en. In: *J. Am. Med. Inform. Assoc.* 21.3 (May 2014), pp. 455–463.

[94] *The Digital Universe: Driving Data Growth in Healthcare.* `https://www.emc.com/analyst-report/digital-universe-healthcare-vertical-report-ar.pdf`. Accessed: 2017-2-23.

[95] Orly Tonkikh et al. "Functional status before and during acute hospitalization and readmission risk identification". en. In: *J. Hosp. Med.* 11.9 (Sept. 2016), pp. 636–641.

[96] Eva Tsui et al. "Development of an automated model to predict the risk of elderly emergency medical admissions within a month following an index hospital visit: A Hong Kong experience". en. In: *Health Informatics J.* (Dec. 2013).

[97] Carl van Walraven et al. "Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community". en. In: *CMAJ* 182.6 (Apr. 2010), pp. 551–557.

[98] Colin Walsh and George Hripcsak. "The effects of data sources, cohort selection, and outcome definition on a predictive model of risk of thirty-day hospital readmissions". en. In: *J. Biomed. Inform.* 52 (Dec. 2014), pp. 418–426.

[99] Daniel A Waxman et al. "A model for troponin I as a quantitative predictor of in-hospital mortality". In: *J. Am. Coll. Cardiol.* 48.9 (2006), pp. 1755–1762.

[100] L L Weed. "Medical records that guide and teach". en. In: *N. Engl. J. Med.* 278.12 (Mar. 1968), 652–7 concl.

[101] Tom L Whitlock et al. "A scoring system to predict readmission of patients with acute pancreatitis to the hospital within thirty days of discharge". en. In: *Clin. Gastroenterol. Hepatol.* 9.2 (Feb. 2011), 175–80, quiz e18.

[102] Yonghui Wu et al. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. In: (Sept. 2016). arXiv: `1609.08144 [cs.CL]`.

[103] Hayato Yamana et al. “Procedure-based severity index for inpatients: development and validation using administrative database”. en. In: *BMC Health Serv. Res.* 15 (July 2015), p. 261.

[104] Antonio Zapatero et al. “Predictive model of readmission to internal medicine wards”. en. In: *Eur. J. Intern. Med.* 23.5 (July 2012), pp. 451–456.

[105] Huaqiong Zhou et al. “Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review”. en. In: *BMJ Open* 6.6 (27 6 2016), e011060.