Title
Predicting Inpatient Medication Orders From Electronic Health Record Data

Journal
Clinical Pharmacology & Therapeutics, 108(1)

ISSN
0009-9236

Authors
Rough, Kathryn
Dai, Andrew M
Zhang, Kun
et al.

Publication Date
2020-07-01

DOI
10.1002/cpt.1826

Peer reviewed

Predicting Inpatient Medication Orders From Electronic Health Record Data

Kathryn Rough[1,*,†], Andrew M. Dai[1], Kun Zhang[1], Yuan Xue[1], Laura M. Vardoulakis[1], Claire Cui[1], Atul J. Butte[2], Michael D. Howell[1] and Alvin Rajkomar[1,3]

In a general inpatient population, we predicted patient-specific medication orders based on structured information in the electronic health record (EHR). Data on over three million medication orders from an academic medical center were used to train two machine-learning models: A deep learning sequence model and a logistic regression model. Both were compared with a baseline that ranked the most frequently ordered medications based on a patient's discharge hospital service and amount of time since admission. Models were trained to predict from 990 possible medications at the time of order entry. Fifty-five percent of medications ordered by physicians were ranked in the sequence model's top-10 predictions (logistic model: 49%) and 75% ranked in the top-25 (logistic model: 69%). Ninety-three percent of the sequence model's top-10 prediction sets contained at least one medication that physicians ordered within the next day. These findings demonstrate that medication orders can be predicted from information present in the EHR.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ Previous medication prediction research has generally been performed within a narrow population, with infrequent timing of predictions (i.e., at the encounter or day level), or by aggregating medications into broad categories.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ This study examines the ability of machine-learning models to provide patient-specific and time-specific predictions of medication orders based on information in the electronic health record.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ This study builds on the clinical applicability of previous work by predicting medication orders in a general inpatient population, whenever orders are placed due to patient needs and clinical workflow, and without aggregating medications into classes.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ Similar models could eventually facilitate patient-specific decision support to reduce the time spent placing orders or assist in the detection of abnormal orders to prevent medication errors. This work represents a first step, and further research in these domains is needed.

The utility of medical predictive models is demonstrated by their ongoing use in clinical care. Routinely used examples include the Pooled Cohort Equation to estimate cardiovascular risk[1] and CHA_2DS_2-VASc to predict thromboembolism.[2] These models produce risk scores for a single outcome that assist clinicians in decision making for a specific question: Is this patient at high enough risk of cardiovascular events to require a statin? Does this patient with atrial fibrillation have sufficiently elevated thromboembolism risk to benefit from anticoagulation? Most well-known clinical predictions come from carefully designed statistical models that use a small set of patient observations and predictor variables; they fall on the simpler end of the "machine-learning spectrum."[3]

Yet, clinical care is not confined to isolated binary decisions; physicians generally choose among many treatment options, and those decisions may change as a patient's condition evolves. For instance, a physician admitting a patient with signs of severe infection may initially order intravenous fluids, an analgesic, and several broad-spectrum antibiotics. Throughout the hospitalization, additional medications will be ordered as adjunctives or to address pre-existing conditions. Eventually, a more targeted antibiotic might be ordered based on blood culture results. Developing predictive models to reliably anticipate these types of heterogeneous therapeutic actions over the course of a hospitalization could lead to increasingly useful clinical decision-support tools.

[1]Google, Mountain View, California, USA; [2]Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, California, USA; [3]Department of Medicine, University of California San Francisco, San Francisco, California, USA. *Correspondence: Kathryn Rough (rough@google.com)

[†]This work was completed as part of the Google AI Residency.

Received October 31, 2019; accepted February 14, 2020. doi:10.1002/cpt.1826

Medication-related prediction and machine-learning research has generally focused on selected patient populations and therapeutic areas. Examples include prediction of prescribing dynamics for sleep medications,[4] onset of vasopressor use in intensive care units,[5] progression through diabetic medications,[6] pharmaceutical treatments in pregnant women,[7] and discharge antihypertensive medications.[8] Numerous methods have also been applied to improving clinical order sets.[9–16] A more general approach was taken by Choi *et al.*[17]; using electronic health record (EHR) data, a recurrent neural network predicted the medication classes likely to be prescribed during the next outpatient encounter.

The clinical applicability of previous work on medication prediction has been limited by narrow patient populations, infrequent timing of predictions (i.e., at the encounter or day level), or by aggregating medications into broad categories. In this paper, we build on previous work by training models to predict which specific medication compounds physicians will order across a general inpatient population throughout their hospitalization, without restricting to particular patient cohorts or therapeutic areas. We evaluate the models' capacity to produce predictions whenever orders are placed due to patient needs and clinical workflow.

METHODS

Study cohort

For all experiments, we used de-identified EHR data from an academic medical center, the University of California San Francisco (UCSF), between 2012 and 2016. We included all adult patients (≥ 18 years of age) who were hospitalized for at least 24 hours. Hospitalizations with no medication orders were excluded. No patients were excluded due to missing or null values. The EHR data contained information from inpatient and outpatient encounters on demographics, diagnoses, procedures, laboratory values, vital signs, flowsheets, and medication orders. Data were de-identified by UCSF before sharing with Google and initiation of analyses.

We partitioned patients into model training (80%), validation (10%), and test sets (10%). Patients in the test set were different from patients in the training set; this means that no medication orders from patients appearing in the training set were included in the test set. To prevent overfitting, the test set remained hidden until the final model evaluation.

De-identification included removal of name, address, phone numbers, email addresses, record and encounter numbers, payer information, physician names, free-text notes, and more, and all dates underwent date-shifting with each patient's dates shifted by a different random number of days up to 1 year (intervals between events were kept consistent within patient records). The EHR data were not joined or combined with any other data. Storage of data was encrypted and access-controlled. Analyses were logged and are auditable. An institutional review board at UCSF issued a research exemption for the de-identified data used in this study.

Data representation

EHR data were structured using an open-source, standardized format for clinical data, the Fast Healthcare Interoperability Resources standard.[18] A detailed description of our data representation and processing approach has been previously published,[19] and the Fast Healthcare Interoperability Resources data representation has been open sourced.[20] For these analyses, events occurring in the EHR were chronologically ordered into a timeline starting from the beginning of a patient's record to the most recent encounter (**Figure 1**). Clinical event time-stamps corresponded to the time of data entry in the EHR. Individual events may be comprised of multiple attributes; for instance, a procedure order could contain a text descriptor of the procedure, the institution-specific procedure code, as well as the equivalent Current Procedural Terminology[21] and Healthcare Common Procedure Coding System[22] codes used for billing.

The data in each event attribute were represented according to their underlying type. Categorical variables (e.g., procedure codes) and text variables (e.g., descriptors of diagnosis codes) were represented as one-hot or multi-hot vectors. Numeric variables (e.g., vital signs) were discretized into deciles, and the deciles were represented as vectors. Embeddings for the vectorized predictors were randomly initialized and jointly trained with the model. Additional data representation details are presented in **Supplementary Note S1** and **Figure S1**.

Description of prediction task

Physicians use computerized order entry systems to write medication orders for inpatients as clinical need arises. As illustrated in **Figure 1**, a prediction was made every time a medication order event occurred in the patient's timeline. A prediction for a medication was considered correct if an order for it was placed within the following 10 minutes. Because multiple medications can be ordered in this timeframe, we consider this prediction task to be multilabel, meaning there can be multiple "correct" medications per time point.[23] Because many medication orders are placed throughout a hospitalization, multiple irregularly spaced predictions were rendered within a single encounter. Predictions were based on the patient's clinical history prior to the order, but no future information.

Medications are often coded according to a hospital-specific formulary that distinguishes between specific brands, manufacturers, and inactive ingredients. Because therapeutically equivalent medications may be coded differently for these administrative purposes, we mapped institution-specific medication codes to RxNorm, a normalized classification system produced by the National Library of Medicine,[24] at the semantic clinical dose form group level. The semantic clinical dose form group unifies medications with the same active ingredients and route of administration (e.g., "morphine injectable product"). This process resulted in 990 distinct medications being used in our analyses.

We compared the performance of a sequence model, a regularized, time-bucketed logistic regression model, and a simple "frequency comparator." The frequency comparator ranked the most frequently ordered medications based on a patient's discharge hospital service and the amount of time since admission. All three methods are described in further detail below.

Sequence model description

For this prediction task, we trained a long short-term memory (LSTM) sequence model,[25] a type of recurrent neural network. Predictors were encoded as vectors and were read by the sequence model in temporal order. An LSTM has an internal state that "remembers" a representation of selected pieces of information it has seen. When predictor information is encountered at a given time step, the LSTM sequence model processes it to determine how to modify its internal state representation (or "memory"): What new information to retain, what old information to forget, and what information to pass along before moving to the next time step. This process is repeated as the LSTM reads through the sequence in temporal order. At the final time step, the LSTM sequence model's internal representation is passed to the network's output layer, and probability estimates are generated for each of the 990 candidate medications.

Embeddings for the predictor vectors were learned during training. Because sequences can be hundreds of thousands of attributes long per patient, predictors were summarized within 12-hour time steps. When multiple observations of the same predictive feature occurred within a time step (e.g., four recorded heart rate measurements), their embeddings were averaged using weights learned during model training.[19] If there were no recorded observations for a predictor in the time step, it was represented as a vector of zeros. Embeddings for all predictors were concatenated, along with an embedding that captured elapsed time between the given time step and the prediction time.

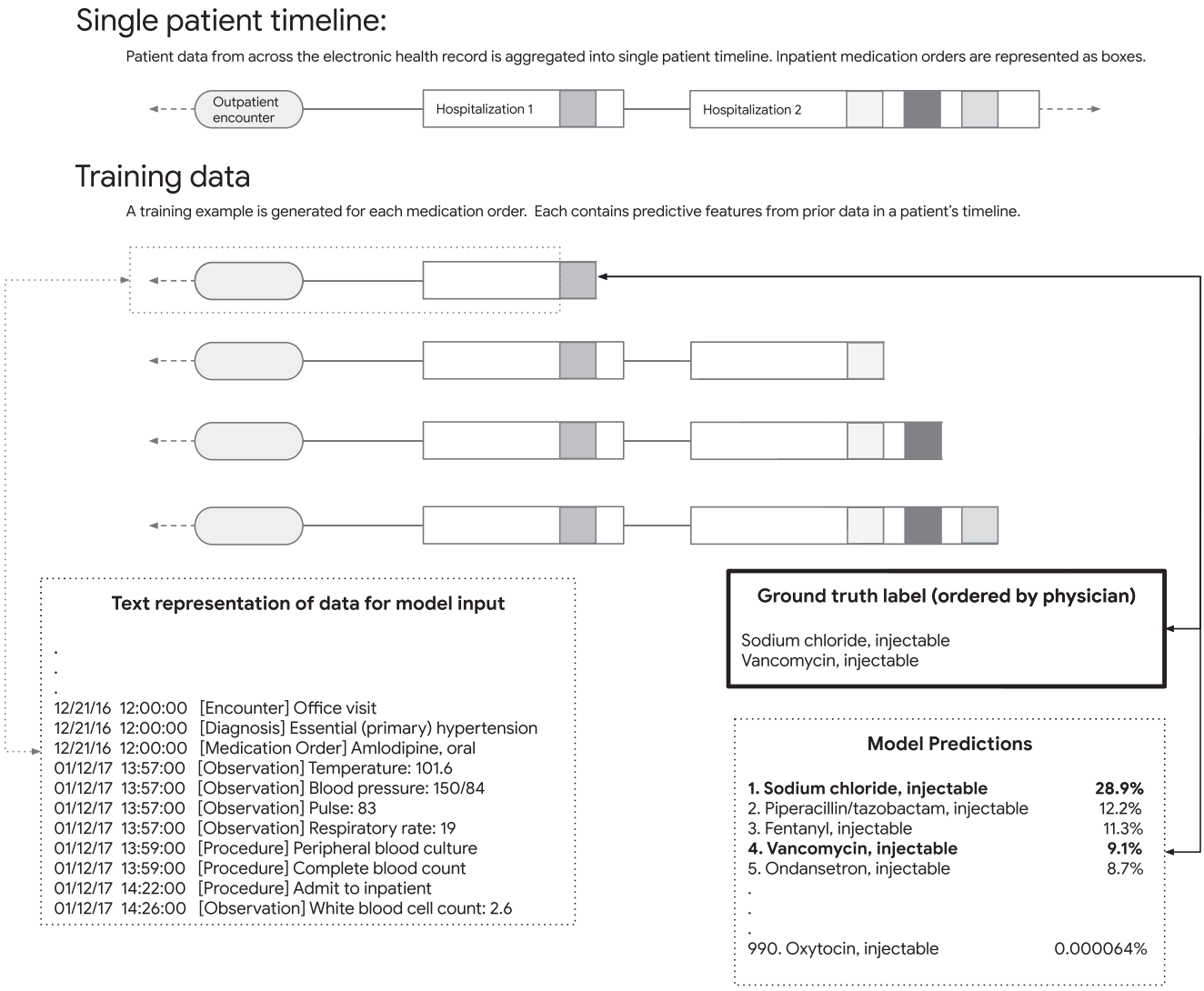

Figure 1 Schematic of study design and prediction task: Illustration of patient timeline, training data, model input, and model predictions for a hypothetical patient. Historical data from a patient's electronic health record is aggregated into a single timeline. In this example, prior outpatient encounters are represented as ovals; inpatient encounters are represented as rectangles. One training example is generated for each inpatient medication order placed (marked by gray squares). Input to the model includes data from the patient's timeline up to the time of the order, but no future information. At the time of each medication order, the model outputs the probability that each of the 990 candidate medications will be ordered within the next 10 minutes. The two drugs ordered by a clinician at this time point (i.e., ground truth) appear in bold type.

All models were trained with TensorFlow in Python.[26,27] LSTM sequence models took ~ 12 days of training time on a single Tesla P100 GPU to converge. The learning rate was manually fine-tuned for this task, but model architecture and other hyperparameter values were adopted from a previous multilabel task (discharge diagnosis prediction) on the same dataset.[19] **Supplementary Note S2** contains a description of LSTM sequence model training and implementation details.

Description of logistic model and frequency comparator

Our implementation of the logistic model is equivalent to training a separate logistic regression classifier for each of the 990 unique medications. Each logistic classifier predicts a single medication, ignoring all other medications.[28] The logistic model was trained on the same set of features as the LSTM model, with one important exception; features were averaged within two buckets: The previous 12 hours and the remainder of the patient history. Although the sequence model can accommodate variable sequence lengths, logistic models require a fixed number of inputs, motivating the averaging procedure within two time periods to retain temporal information. Parameters were not shared between individual classifiers, resulting in a model with over 200 million trainable parameters. For variable selection and prevention of overfitting, L1 regularization was used. The logistic model trained to convergence in 3 weeks on 10 Tesla P100 GPUs. The learning rate and L1 regularization parameter were manually tuned for this task.

The frequency comparator predicts the most frequently ordered medications based on the patient's discharge hospital service and the time since admission when the medication order was placed. More specifically, the frequency comparator ranks the top-k medications stratified by both the time of order placement (< 1 day, 1–3 days, 3–5 days, and > 5 days) and hospital discharge service (27 total services; see **Table 1**). Experiments showed slightly improved performance using a combination of time since admission and hospital discharge service, compared with either of these

Table 1 Descriptive characteristics in training, validation, and test sets for inpatient medication order prediction task

	Training set	Validation set	Test set
Patient-level characteristics	**(*N* = 63,601)**	**(*N* = 6,504)**	**(*N* = 6,383)**
Sex, *n* %			
Female	35,957 (56.5)	3,607 (55.5)	3,680 (57.7)
Male	27,638 (43.5)	2,895 (44.5)	2,702 (42.3)
Unknown	6 (< 0.1)	2 (< 0.1)	1 (< 0.1)
Race, *n* %			
White	35,791 (56.3)	3,598 (55.3)	3,658 (57.3)
Black/African American	4,913 (7.7)	500 (7.7)	485 (7.6)
Asian/Pacific Islander	9,786 (15.4)	1,024 (15.7)	962 (15.1)
Other (including multiple races)	10,456 (16.4)	1,118 (17.2)	1,055 (16.5)
Unknown	2,655 (4.2)	264 (4.1)	223 (3.5)
Encounter-level characteristics	**(*N* = 98,479)**	**(*N* = 9,907)**	**(*N* = 9,978)**
Age, *n* %			
18–34 years	19,326 (19.6)	2,093 (20.6)	2,033 (20.4)
35–64 years	49,029 (49.8)	4,940 (49.9)	4,960 (49.7)
65–85 years	25,618 (26.0)	2,482 (25.1)	2,561 (25.7)
>85 years	4,506 (4.6)	446 (4.5)	424 (4.2)
Hospital discharge service, *n* (%)			
General medicine	21,803 (21.1)	2,129 (21.5)	2,168 (21.7)
Neurosurgery	10,465 (10.6)	1,068 (10.8)	1,047 (10.5)
Obstetrics	10,325 (10.5)	1,071 (10.8)	1,064 (10.7)
Orthopedics	7,741 (7.9)	762 (7.7)	839 (8.4)
Transplant	7,393 (7.5)	753 (7.6)	707 (7.1)
General surgery	6,616 (6.7)	628 (6.3)	712 (7.1)
Cardiology	5,770 (5.9)	551 (5.6)	606 (6.1)
Oncology	5,260 (5.3)	575 (5.8)	515 (5.2)
Urology	3,745 (3.8)	384 (3.9)	353 (3.5)
All other services[a]	19,361 (19.7)	1,986 (20.0)	1,967 (19.7)
Previous hospitalizations, *n* (%)			
None	63,284 (64.2)	6,478 (65.4)	6,346 (63.6)
One	16,709 (17.0)	1,702 (17.2)	1,686 (16.9)
Two or more	18,522 (18.8)	1,727 (17.4)	1,946 (19.5)
Medication order event characteristics	**(*N* = 2,433,397)**	**(*N* = 300,176)**	**(*N* = 297,615)**
Number of medications per order event, median (25th percentile, 75th percentile)	3 (2, 5)	2 (2, 4)	2 (2, 4)
Medications ordered	**(*N* = 5,521,361)**	**(*N* = 681,896)**	**(*N* = 685,638)**

[a]Other services: Cardiac surgery, colorectal surgery, critical care medicine, emergency medicine, gynecological oncology, gynecology, hepatobiliary medicine, medical speciality, neurology, oral/maxillofacial surgery, other surgery, otorhinolaryngology, pediatric (≥ 18 years of age), plastic surgery, pulmonary medicine, thoracic surgery, vascular surgery, and other/unknown.

categorizations alone. **Table S1** displays the top medications according to the frequency comparator for several discharge services.

Evaluation of model performance

We computed several metrics of model performance in the held-out test set: Top-k recall, micro-weighted area under the precision recall curve, and micro-weighted area under the receiver operating curve. Top-k recall captures the proportion of medications ordered by physicians that appeared in the model's top-5, top-10, top-15, and top-25 most probable predictions (recall is synonymous with sensitivity). Top-10 recall was additionally reported stratified by hospital discharge service. Alternative calculations of the top-k recall metrics (where true medication labels do not count towards k) are presented in **Table S2**.

The area under the precision recall curve (AU-PRC) contrasts the tradeoffs between recall and precision over a range of thresholds (precision is synonymous with positive predictive value). The AU-PRC is generally considered a more informative alternative to the area under the receiver operating curve (AU-ROC) when the distribution of outcome labels is highly

skewed.[29] The micro-weighted AU-ROC measures model discrimination—the trade-off between recall and false-positive rate—over a range of cutoff thresholds. Both the AU-PRC and AU-ROC require micro-weighting[30] to generalize metrics developed for binary predictions to the multilabel setting (metrics are aggregated over medication-specific contingency tables). We used 1,000 bootstrapped samples to calculate 95% confidence intervals around all metrics.

To evaluate whether the model's output anticipated orders placed within a short timeframe, we calculated two forms of a top-k precision metric. The first is the proportion of the top model-generated predictions that were ordered by a physician within a time window. The second is the proportion of top-k model-generated prediction sets where at least one medication was ordered within a time window. Both precision metrics were calculated at 30-minute intervals from the time of the prediction to 24 hours after the prediction; patients discharged during the time window were not censored (i.e., they were retained in the denominator). This calculation relies on the assumption that patients do not receive orders for additional medications during a short period immediately postdischarge; violations of this assumption would result in underestimated metrics.

All model evaluation was performed using Python.

RESULTS

Characterizing the patient population

There were 76,488 unique patients with 118,364 eligible hospitalizations, resulting in 6,888,895 individual medications from 3,031,188 distinct order events. The training set contained 2.4 million medication order events from 98,479 encounters and 63,601 unique patients (**Table 1**). There were more patients identifying as women than men (57% vs. 44%). Fifty-six percent of patients identified as white, 15% as Asian/Pacific Islander, and 8% as black/African American. Half of the encounters were for patients 35–64 years old, and 64% were the first-recorded hospitalizations for patients. Patients were discharged from a range of hospital services—general medicine (21%), neurosurgery (11%), and obstetrics (11%) were most common. Descriptive characteristics seemed to be similar across the training, validation, and test sets (**Table 1**).

Medication prediction task: Overall results

Admissions had a median of 19 distinct medication order events (25th percentile: 12 and 75th percentile: 33); a prediction was made each time an order was placed (**Figure 1**).

The held-out test set contained 298,000 order events. In it, the sequence model had a top-10 recall of 55%, meaning over half of all medications ordered by physicians were ranked in the top-10 (out of nearly 1,000 possibilities; **Table 2**). To ensure the model's performance was not explained by memorizing and predicting previous medications ordered for the patient, we trained a sequence model that provided no information on previous medication-related predictors. This resulted in modest reductions in recall (top-10: 52%). The frequency comparator had substantially lower recall for all cutoffs (top-10 recall: 32%), and the logistic model's performance fell between that of the sequence and the frequency comparator (top-10 recall: 49%). The median predicted rank of medications actually ordered by physicians was 9 for the sequence model (25th percentile: 3 and 75th percentile: 26), 11 for the logistic model (25th percentile: 4 and 75th percentile: 36), and 23 for the frequency comparator (25th percentile: 7 and 75th percentile: 67).

One limitation of the recall metric is that correct predictions can be "crowded out" of the top-k when multiple medications are ordered simultaneously or in quick succession. An alternative calculation, where correctly predicted medications are not counted toward the top-k, increases the top-10 recall to 62% for the sequence model and 52% for the logistic model (**Table S2**).

The micro-weighted AU-PRC was 0.299 (95% confidence interval 0.297–0.300) for the sequence model and 0.193 for the logistic model (95% confidence interval 0.191–0.195). The five individual medications with the highest AU-PRCs were injectable carboprost, injectable methylergonovine, injectable oxytocin, oral vardenafil, and injectable terbutaline. The five individual medications with the lowest AU-PRCs were oral liquid pyridostigmine, oral liquid entecavir, oral cevimeline, oral cefuroxime, and ophthalmic levobunolol.

Illustrative prediction example

Figure 2 illustrates medication predictions generated by the sequence model at three time points during a patient's week-long hospitalization. The patient had a history of kidney transplantation and chronic hypertension and presented to care with shortness of breath and symptoms consistent with a respiratory infection. The top-10

Table 2 Model performance for inpatient medication prediction task, with 95% confidence intervals[a] (measured in the held-out test set)

	LSTM sequence model: All variables	LSTM sequence model: No medication variables[b]	Logistic regression model	Frequency comparator[c]
Top-5 recall[d]	0.390 (0.389, 0.391)	0.360 (0.359, 0.390)	0.340 (0.339, 0.342)	0.209 (0.207, 0.212)
Top-10 recall[d]	0.552 (0.550, 0.553)	0.520 (0.519, 0.552)	0.489 (0.487, 0.490)	0.324 (0.320, 0.327)
Top-15 recall[d]	0.645 (0.644, 0.646)	0.615 (0.614, 0.644)	0.579 (0.577, 0.581)	0.408 (0.404, 0.413)
Top-25 recall[d]	0.750 (0.749, 0.751)	0.723 (0.722, 0.745)	0.686 (0.684, 0.687)	0.527 (0.523, 0.530)
Micro-weighted AU-PRC	0.299 (0.297, 0.300)	0.258 (0.257, 0.299)	0.193 (0.191, 0.195)	—
Micro-weighted AU-ROC	0.977 (0.977, 0.977)	0.974 (0.974, 0.977)	0.956 (0.955, 0.956)	—

Many medications can be ordered simultaneously in inpatient settings: 20% of order events have ≥ 5 medications actually ordered, 7% have ≥ 10 medications ordered, 4% have ≥ 15 medications ordered, 1% have ≥ 25 medications ordered. This places a bound on the highest attainable top-k recall metrics (i.e., a perfect model would achieve only 80% top-5 recall). For alternative calculation of recall metrics, please see **Table S1**.

AU-PRC, area under the precision-recall curve; AU-ROC, area under the receiver operating curve; LSTM, long short-term memory.

[a]95% Confidence intervals were calculated from 1,000 bootstrapped samples. [b]LSTM sequence model trained with no information on previous medications. [c]The frequency comparator ranked the top-k most frequently ordered medications based on the time between hospital admission and the placement of the order (< 1 day, 1–3 days, 3–5 days, and > 5 days) and the patient's discharge hospital service. [d]Top-k recall is the proportion of medications actually ordered by physicians that appear in the model's top-k most probable medication predictions.

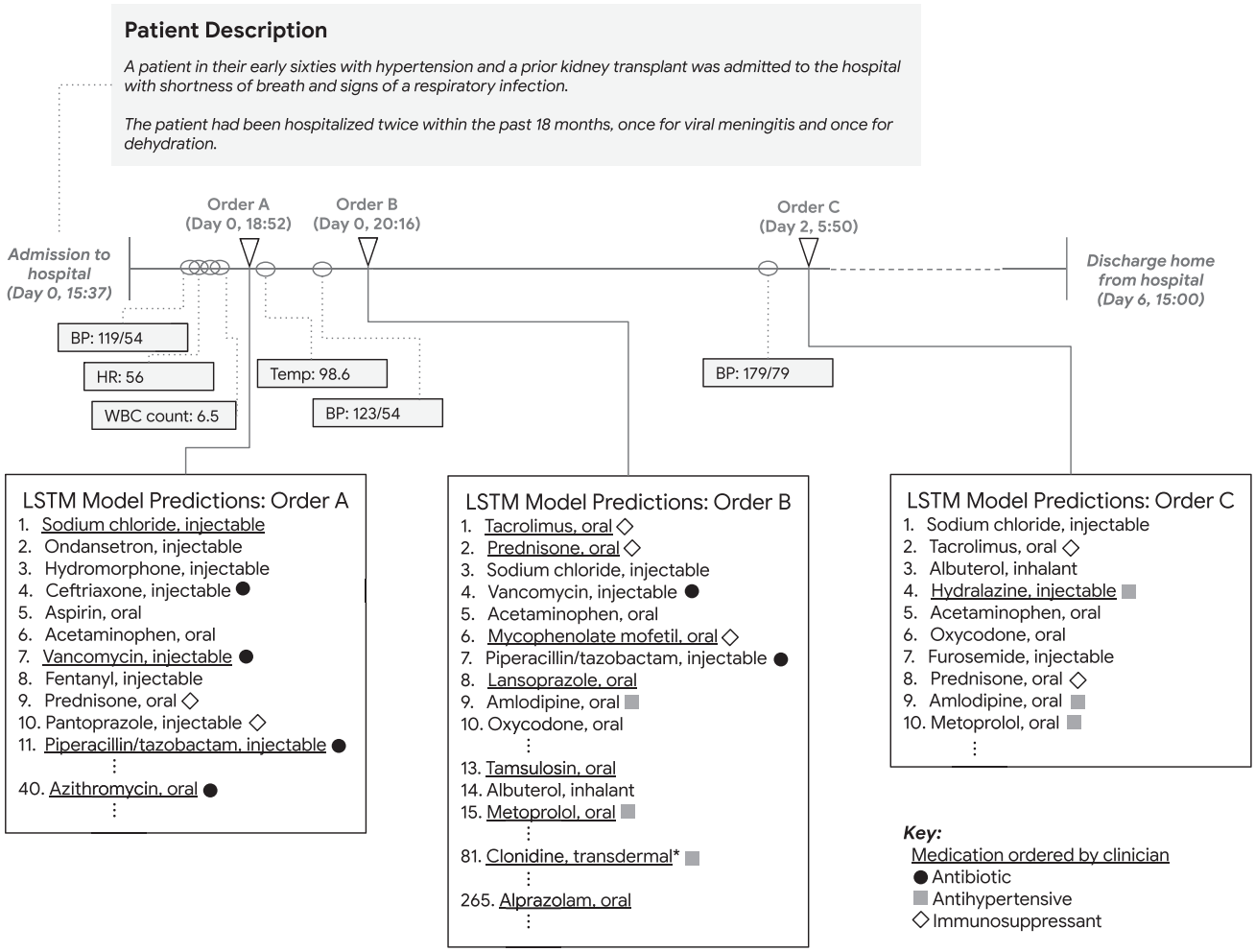

Figure 2 Illustrative example of LSTM sequence model-generated medication predictions for a single patient. Predictions produced by the LSTM sequence model at three time points during a patient's hospitalization are shown. Medications that were actually placed by clinicians (i.e., ground truth) are underlined. We display a small subset of vital signs and laboratory results collected during the hospitalization, represented as empty circles. Black circles denote antibiotic medications, gray squares denote antihypertensive medications, and white diamonds denote immunosuppressant medications. Near admission, the model predicts multiple antibiotics and pain relievers (order A). Soon after, for order B, the model assigns high probability for multiple immunosuppressive medications typically administered to transplant patients. Several days into the hospitalization, the patient's blood pressure rises; the LSTM sequence model predicts a variety of antihypertensives, including an intravenous formulation of hydralazine (order C). BP, blood pressure (measured in millimeters of mercury); HR, heart rate (measured in beats per minute); LSTM, long short-term memory; Temp, temperature (measured in degrees Fahrenheit); WBC, white blood cell (measured in thousands of cells per microliter). *"Clonidine, oral" was ranked 36 by the LSTM sequence model at this time point. Note: Corresponding predictions for the logistic model can be found in **Figure S2**.

predictions made at the first time point (order A) included medications commonly ordered for patients at the time of admission: Broad-spectrum antibiotics, treatments used for rehydration, pain relievers, and antinausea medications. Three of four medications ordered for the patient appear in the model's top-15 predictions.

Order B was placed later that evening. The model's top-10 predictions included medications more specific to the patient's needs, including immunosuppressants and antihypertensives. Of the 8 medications ordered, 6 appeared in the top-15; however, the model assigned a low probability to the transdermal clonidine and oral alprazolam ordered by physicians (ranked 81 and 256).

Two days later, the patient's blood pressure rose acutely, despite oral antihypertensive use throughout the hospitalization. An order was placed for injectable hydralazine, an antihypertensive used for short-term blood pressure control. The medication was correctly ranked in the top-10, as were two additional antihypertensives. Predictions generated by the logistic model appear in **Figure S2**.

Model performance results stratified by hospital discharge service

Stratification by hospital discharge service reveals heterogeneity in model performance (**Figure 3**). Some services treat patients with relatively homogenous needs, leading to a narrower set of medications typically ordered; the models tended to have better performance in these settings compared with services that care for patients with more homogeneous conditions that require

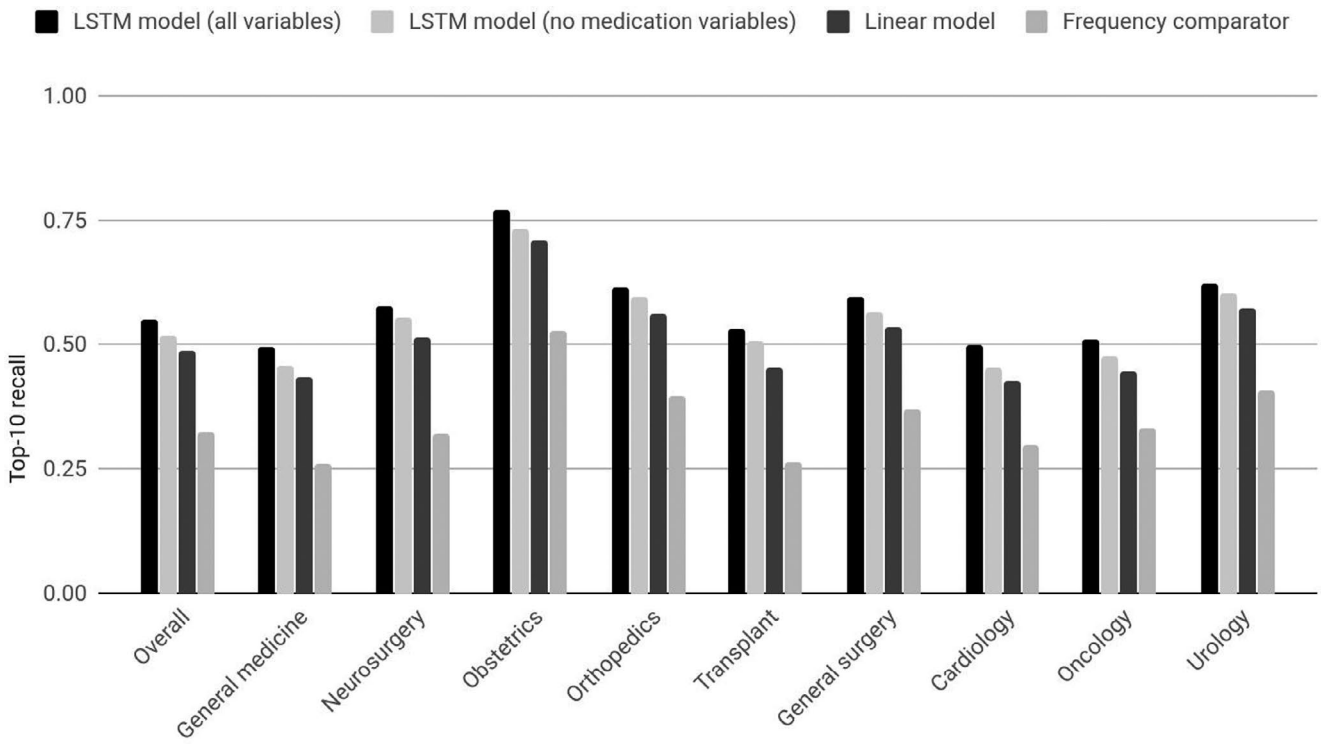

Figure 3 Top-10 recall for the inpatient medication prediction task by hospital discharge service (measured in the held-out test set). Top-10 recall is the proportion of medications actually ordered by physicians that appear in the model's top-10 most probable medication predictions. LSTM, long short-term memory.

a wider set of medications. For instance, the sequence model's recall is highest for patients discharged from obstetrics (top-10 recall: 77%), where 80% of medication orders are covered by 27 medications. Sequence model recall is lower for patients discharged from cardiology (top-10 recall: 50%) and general medicine (top-10 recall: 50%), where 89 and 109 unique medications, respectively, cover 80% of medication orders. Similar performance patterns were observed for the logistic model and frequency comparator.

Precision of predictions over time

Figure 4a provides insight into whether the sequence model's output anticipates subsequent orders placed by physicians in the near future. Within 12 hours, 51% of the model's top-1 predictions, 36% of the model's top-5 predictions, and 29% of the model's top-10 predictions were ordered (**Figure 4a**). Within 24 hours, the proportions increased to 60% for top-1 predictions, 44% for top-5 predictions, and 35% for top-10 predictions. (**Figure S3** displays results for the logistic model.)

Figure 4b illustrates complementary information; the proportion of top-k sequence model-predicted sets where at least one medication was ordered within a 24-hour timeframe. Within 12 hours, 78% of the model's top-5 sets had at least one medication ordered and 87% of the model's top-10 sets had at least one medication ordered. Within 24 hours, these numbers increased to 87% for top-5 prediction sets and 93% for top-10 prediction sets. (**Figure S4** displays results for the logistic model.)

Additional analyses

Several additional analyses were performed to better understand the sequence model's false positive and false negative predictions using a previously published taxonomy of medication classes based on therapeutic area[31] (**Supplementary Note S3**; **Tables S4** and **S5**). Experiments involving predictive feature ablation are also presented in the **Supplementary Materials** (**Supplementary Note S4**; **Figures S5**and **S6**).

DISCUSSION

Using EHR data, machine-learning models can be trained to provide patient-specific and time-specific predictions of medication orders for hospitalized patients. From 990 possible medications, 55% of medications ordered by physicians appeared in the sequence model's top-10 predictions at the time of order placement. Nearly all (93%) top-10 prediction sets contained at least one medication that would be ordered by clinicians within the next day. This performance was not explained by the model simply predicting previously ordered medications or by repeatedly predicting medications commonly ordered for patients in the same hospital service. The sequence model's performance exceeded that of a regularized, time-bucketed logistic regression model. We also found that the medication prediction task was not uniformly difficult across all types of patients; all models performed better in groups with narrower use of medications compared with those with more heterogeneous therapeutic needs.

Our findings indicate that it is feasible to use machine learning to predict patterns of physician medication ordering despite the

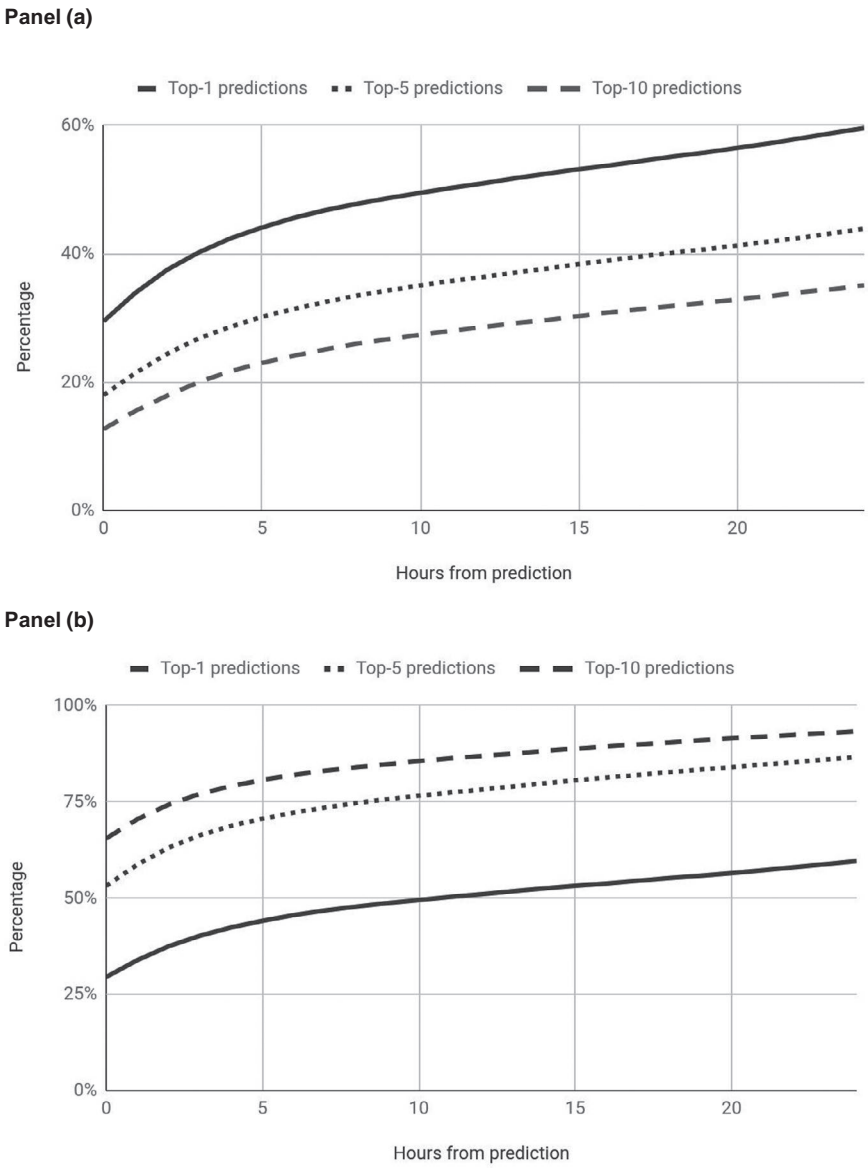

Figure 4 Physician ordering of top LSTM sequence model-predicted medications within a specified postprediction time window (measured in the held-out test set). (**a**) The percentage of top-1, top-5, and top-10 LSTM sequence model-predicted medications ordered by physicians within the specified postprediction time window. (**b**) The percentage of top-1, top-5, and top-10 LSTM sequence model-predicted sets where at least one medication is ordered by physicians within the specified postprediction time window. LSTM, long short-term memory.

multiple nuanced factors in the prediction. Predicting physician medication orders for inpatients includes a broad set of potential medications, any number of which can be ordered in close proximity. The precise temporal sequencing of orders is somewhat arbitrary, and many aspects of the clinical workflow that influence timing of orders will not be captured in EHR data. There is also substantial variability in physician prescribing,[32–34] to the extent that individual physician prescribing preferences have been used as an instrumental variable in studies of medication safety and effectiveness.[35–37]

Model predictions should not be interpreted as optimal treatment; using observational data to estimate the comparative safety or effectiveness of medications requires a causal inference framework.[38,39] Instead, models were trained to reproduce physician behavior as it appears in historical data, which may not be consistent with current clinical guidelines. However, previous work has shown that training recommender system algorithms based on historical hospital ordering data resulted in order sets more aligned with practice guidelines than manually authored hospital order sets.[40]

Despite being outcome-agnostic, inpatient medication predictions may have useful clinical decision support applications. Inpatient medication errors cause substantial morbidity and mortality, occurring in an estimated 3.8 million hospitalizations each year.[41–44] There is some evidence that prediction of typical clinician decision making may allow for patient-specific anomaly detection.[45,46] In future work, we intend to investigate whether our model could facilitate improved real-time detection and alerting for medication errors.

This study has several important limitations. First, this work used data from a single site, limiting generalizability. Due to

differences in patient populations, prescribing practices, and storage of health information in EHRs, the model would likely need to be retrained for use in different settings. However, our approach is general; we use a data representation, predictive features, and model architecture that are not hospital-specific or task-specific.[19] Second, only retrospective data were used to evaluate the model. Due to temporal changes in prescribing practices, including the addition of new medications to formularies, it is likely that models would need to be regularly retrained with new data to prevent degradation of model performance. Third, our study focused on the prediction of medication compounds, specifically excluding dosages. Fourth, there is evidence of racial and gendered inequities in prescribing for some conditions, particularly pain management[47–49]; the data used to train these models may also contain these prescribing biases. However, multiple approaches may help to address some of these biases.[50] Fifth, the current models were trained and evaluated conditional on a medication order being placed; a separate mechanism would be necessary to know when to run the model in a prospective deployment. One reasonable possibility would be generating a prediction whenever clinicians navigate to the EHR's order-entry screen. Finally, sequence models are often viewed as difficult to interpret. However, our previous work has demonstrated that attribution methods can provide useful insight into which data elements influence a specific prediction.[19]

Our study also has several strengths. We demonstrate that machine-learning models can be used to predict the inpatient medication orders placed by clinicians. Our approach was flexible and general; we included all adult inpatients, without restricting to disease-related subgroups or relying on hand-curated predictors. The model was capable of predicting individual medication compounds, not only medication classes. Predictions can be generated at a clinically relevant time point, based on all available information at the time of entry into the medication ordering system.

These findings represent an incremental step forward in this domain; we anticipate and encourage future research that will improve upon this initial approach to increase predictive performance, interpretability, fairness, or generalizability of the model across clinical institutions. It remains to be seen whether medication order prediction will facilitate better clinical decision support systems.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutics* website (www.cpt-journal.com).

ACKNOWLEDGMENTS

The authors thank members of our broader research team who have assisted this project through the development of analytical tools, data collection, maintenance of research infrastructure, assurance of data quality, and project management, including Gabby Espinosa, Gerardo Flores, Michaela Hardt, Sharat Israni, Hong Ji, Jeff Love, Dana Ludwig, Svetlana Kelman, I-Ching Lee, Mimi Sun, Patrik Sundberg, Chunfeng Wen, and Doris Wong. For their helpful comments and revisions to this manuscript, we would like to recognize Gerardo Flores, Michaela Hardt, Yun Liu, Stelios Serghiou, Ethan Steinberg, and Zhen Xu.

FUNDING

Research reported in this publication was supported in part by funding from the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1 TR001872. Google additionally provided funding for the work.

CONFLICT OF INTEREST

Google employs K.R., A.M.D., K.Z., Y.X., L.M.V., C.C., M.D.H., and A.R., who also own equity in the company. A.J.B. is cofounder and consultant to Personalis and NuMedii; consultant to Samsung, Geisinger Health, Mango Tree Corporation, Regenstrief Institute, and in the recent past, 10x Genomics and Helix; shareholder in Personalis; minor shareholder in Apple, Facebook, Google, Microsoft, Sarepta, 10x Genomics, Amazon, Biogen, CVS, Illumina, Snap, and Sutro, and several other nonhealth-related companies and mutual funds; and has received honoraria and travel reimbursement for invited talks from Genentech, Roche, Pfizer, Merck, Lilly, Mars, Siemens, Optum, AbbVie, Westat, and many academic institutions, medical or disease specific foundations and associations, and health systems. A.J.B. receives royalty payments through Stanford University, for several patents and other disclosures licensed to NuMedii and Personalis. A.J.B.'s research has been funded by the National Institutes of Health (NIH), Northrup Grumman (as the prime on an NIH contract), Genentech, US Food and Drug Administration (FDA), the Leon Lowenstein Foundation, the Intervalien Foundation, Priscilla Chan and Mark Zuckerberg, the Barbara and Gerson Bakar Foundation, and, in the recent past, the March of Dimes, Juvenile Diabetes Research Foundation, California Governor's Office of Planning and Research, California Institute for Regenerative Medicine, L'Oreal, and Progenity.

AUTHOR CONTRIBUTIONS

All authors wrote the manuscript. K.R., A.M.D., M.D.H., C.C., A.J.B., and A.R. designed the research. K.R. and A.M.D. performed the research and analyzed the data. K.R., A.M.D., K.Z., and Y.X. contributed to new analytical tools.

DATA AVAILABILITY STATEMENT

The datasets analyzed in this study are not publicly available. Due to patient privacy and security concerns, the underlying EHR data are not easily redistributable to researchers other than those engaged in institutional review board–approved collaborations with the University of California San Francisco.

1. Muntner, P. *et al*. Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. *JAMA* **311**, 1406–1415 (2014).
2. Knaus, W.A. *et al*. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* **100**, 1619–1636 (1991).
3. Beam, A.L. & Kohane, I.S. Big data and machine learning in health care. *JAMA* **319**, 1317–1318 (2018).
4. Beam, A.L. *et al*. Predictive modeling of physician-patient dynamics that influence sleep medication prescriptions and clinical decision-making. *Sci. Rep.* **7**, 42282 (2017).
5. Ghassemi, M., Wu, M., Hughes, M.C., Szolovits, P. & Doshi-Velez, F. Predicting intervention onset in the ICU with switching state space models. *AMIA Jt Summits Transl. Sci. Proc.* **2017**, 82–91 (2017).
6. Wright, A.P., Wright, A.T., McCoy, A.B. & Sittig, D.F. The use of sequential pattern mining to predict next prescribed medications. *J. Biomed. Inform.* **53**, 73–80 (2015).
7. Klann, J., Schadow, G. & Downs, S.M. A method to compute treatment suggestions from local order entry data. *AMIA Annu. Symp. Proc.* **2010**, 387–391 (2010).
8. Yang, Y. *et al*. Predicting discharge medications at admission time based on deep learning. *arXiv [cs.CL]* (2017) <http://arxiv.org/abs/1711.01386>.

9. Gartner, D., Zhang, Y. & Padman, R. Cognitive workload reduction in hospital information systems: decision support for order set optimization. *Health Care Manag. Sci.* **21**, 224–243 (2018).
10. Klann, J., Schadow, G. & McCoy, J.M. A recommendation algorithm for automating corollary order generation. *AMIA Annu. Symp. Proc.* **2009**, 333–337 (2009).
11. Klann, J.G., Szolovits, P., Downs, S.M. & Schadow, G. Decision support from local data: creating adaptive order menus from past clinician behavior. *J. Biomed. Inform.* **48**, 84–93 (2014).
12. Zhang, Y., Padman, R. & Levin, J.E. Paving the COWpath: data-driven design of pediatric order sets. *J. Am. Med. Inform. Assoc.* **21**, e304–e311 (2014).
13. Chen, J.H., Podchiyska, T. & Altman, R.B. OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records. *J. Am. Med. Inform. Assoc.* **23**, 339–348 (2016).
14. Wright, A. & Sittig, D.F. Automated development of order sets and corollary orders by data mining in an ambulatory computerized physician order entry system. In *AMIA Annual Symposium Proceedings Archive*, 819–823 (American Medical Informatics Association, Bethesda, MD, 2006).
15. Chen, J.H., Alagappan, M., Goldstein, M.K., Asch, S.M. & Altman, R.B. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. *Int. J. Med. Inform.* **102**, 71–79 (2017).
16. Chen, J.H., Goldstein, M.K., Asch, S.M., Mackey, L. & Altman, R.B. Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. *J. Am. Med. Inform. Assoc.* **24**, 472–480 (2017).
17. Choi, E., Bahadori, M.T., Schuetz, A., Stewart, W.F. & Sun, J. Doctor AI: predicting clinical events via recurrent neural networks. *JMLR Workshop Conf. Proc.* **56**, 301–318 (2016).
18. Fast Healthcare Interoperability Resources (FHIR) Release 3 (STU; v3.0.1-11917) <https://www.hl7.org/fhir/overview.html>.
19. Rajkomar, A. *et al*. Scalable and accurate deep learning with electronic health records. *NPJ Digit. Med.* **1**, 18 (2018).
20. Google Fast Healthcare Interoperability Resources (FHIR) protocol buffers <https://github.com/google/fhir>.
21. American Medical Association CPT® (Current Procedural Terminology) <https://www.ama-assn.org/practice-management/cpt-current-procedural-terminology>.
22. Centers for Medicare and Medicaid Services HCPCS General Information <https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/index.html> (2018).
23. Zhang, M.L. & Zhou, Z.H. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.* **26**, 1819–1837 (2014).
24. Unified Medical Language System RxNorm Technical Documentation <https://www.nlm.nih.gov/research/umls/rxnorm/docs/index.html> (2018).
25. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
26. Abadi, M.*et al*. TensorFlow: large-scale machine learning on heterogeneous distributed systems. *arXiv [cs.DC]* (2016) <http://arxiv.org/abs/1603.04467>.
27. TensorFlow <https://github.com/tensorflow/tensorflow>.
28. Tsoumakas, G. & Katakis, I. Multi-label classification: an overview. *Int. J. Data Warehousing and Mining*, **2007**, 1–2 (2007).
29. Davis, J. & Goadrich, M. The relationship between precision-recall and ROC Curves. In *Proceedings of the 23rd International Conference on Machine Learning*, 233–240 (Association for Computing Machinery, New York, NY, 2006).
30. Sebastiani, F. Machine learning in automated text categorization. *ACM Comput. Surv.* **34**, 1–47 (2002).
31. Shehab, N. *et al*. US Emergency Department visits for outpatient adverse drug events, 2013–2014. *JAMA* **316**, 2115–2125 (2016).
32. Carrin, G. Drug prescribing: a discussion of its variability and (ir) rationality. *Health Policy* **7**, 73–94 (1987).
33. Zhang, Y., Baicker, K. & Newhouse, J.P. Geographic variation in Medicare drug spending. *N. Engl. J. Med.* **363**, 405–409 (2010).
34. Zhang, Y., Baicker, K. & Newhouse, J.P. Geographic variation in the quality of prescribing. *N. Engl. J. Med.* **363**, 1985–1988 (2010).
35. Brookhart, M.A., Wang, P.S., Solomon, D.H. & Schneeweiss, S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* **17**, 268–275 (2006).
36. Rassen, J.A., Brookhart, M.A., Glynn, R.J., Mittleman, M.A. & Schneeweiss, S. Instrumental variables II: instrumental variable application-in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance. *J. Clin. Epidemiol.* **62**, 1233–1241 (2009).
37. Brookhart, M.A. & Schneeweiss, S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *Int. J. Biostat.* **3**, Article 14 (2007).
38. Pearl, J., Glymour, M. & Jewell, N.P. *Causal Inference in Statistics: A Primer* (John Wiley & Sons, Chichester, UK, 2016) <https://market.android.com/details?id=book-L3G-CgAAQBAJ>.
39. Hernan, M.A. & Robins, J.M. *Causal Inference* (Chapman & Hall/CRC, Boca Raton, FL, 2018) <https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/>.
40. Wang, J.K. *et al*. An evaluation of clinical order patterns machine-earned from clinician cohorts stratified by patient mortality outcomes. *J. Biomed. Inform.* **86**, 109–119 (2018).
41. Institute of Medicine. *To Err Is Human: Building a Safer Health System* (The National Academies Press, Washington, DC, 2000).
42. Preventing medication errors: a $21 billion opportunity. AHRQ Patient Safety Network <https://psnet.ahrq.gov/resources/resource/20529>.
43. Network for Excellence in Health Innovation. Saving lives, saving money: the imperative for computerized physician order entry in Massachusetts hospitals <https://www.nehi.net/publications/39-saving-lives-saving-money-the-imperative-for-computerized-physician-order-entry-in-massachusetts-hospitals/view>.
44. James, J.T. A new, evidence-based estimate of patient harms associated with hospital care. *J. Patient Saf.* **9**, 122–128 (2013).
45. Hauskrecht, M. *et al*. Conditional outlier detection for clinical alerting. *AMIA Annu. Symp. Proc.* **2010**, 286–290 (2010).
46. Schiff, G.D. *et al*. Screening for medication errors using an outlier detection system. *J. Am. Med. Inform. Assoc.* **24**, 281–287 (2017).
47. Sabin, J.A. & Greenwald, A.G. The influence of implicit bias on treatment recommendations for 4 common pediatric conditions: pain, urinary tract infection, attention deficit hyperactivity disorder, and asthma. *Am. J. Public Health* **102**, 988–995 (2012).
48. Pletcher, M.J., Kertesz, S.G., Kohn, M.A. & Gonzales, R. Trends in opioid prescribing by race/ethnicity for patients seeking care in US emergency departments. *JAMA* **299**, 70–78 (2008).
49. Raftery, K.A., Smith-Coggins, R. & Chen, A.H. Gender-associated differences in emergency department pain management. *Ann. Emerg. Med.* **26**, 414–421 (1995).
50. Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G. & Chin, M.H. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.* **169**, 866–872 (2018).