Journal of the American Medical Informatics Association, 26(12), 2019, 1427–1436
doi: 10.1093/jamia/ocz171
Advance Access Publication Date: 3 October 2019
Research and Applications

Research and Applications

Predicting emergency department orders with multilabel machine learning techniques and simulating effects on length of stay

Haley S. Hunter-Zinck [1], Jordan S. Peck,[2,3] Tania D. Strout,[3,4] and Stephan A. Gaehde[1,5]

[1]Department of Emergency Services, VA Boston Healthcare System, Boston, Massachusetts, USA, [2]Center for Performance Improvement, MaineHealth, Portland, Maine, USA, [3]Department of Emergency Medicine, Tufts University School of Medicine, Medford, Massachusetts, USA, [4]Department of Emergency Medicine, Maine Medical Center, Portland, Maine, USA and [5]School of Medicine, Boston University, Boston, Massachusetts, USA

Corresponding Author: Haley S. Hunter-Zinck, PhD, Department of Emergency Services, VA Boston Healthcare System, 1400 VFW Parkway, West Roxbury, MA 02132, USA; haley.hunter-zinck@va.gov

Received 11 March 2019; Revised 26 August 2019; Editorial Decision 28 August 2019; Accepted 30 August 2019

ABSTRACT

Objective: Emergency departments (EDs) continue to pursue optimal patient flow without sacrificing quality of care. The speed with which a healthcare provider receives pertinent information, such as results from clinical orders, can impact flow. We seek to determine if clinical ordering behavior can be predicted at triage during an ED visit.

Materials and Methods: Using data available during triage, we trained multilabel machine learning classifiers to predict clinical orders placed during an ED visit. We benchmarked 4 classifiers with 2 multilabel learning frameworks that predict orders independently (binary relevance) or simultaneously (random *k*-labelsets). We evaluated algorithm performance, calculated variable importance, and conducted a simple simulation study to examine the effects of algorithm implementation on length of stay and cost.

Results: Aggregate performance across orders was highest when predicting orders independently with a multilayer perceptron (median F_1 score = 0.56), but prediction frameworks that simultaneously predict orders for a visit enhanced predictive performance for correlated orders. Visit acuity was the most important predictor for most orders. Simulation results indicated that direct implementation of the model would increase ordering costs (from $21 to $45 per visit) but reduce length of stay (from 158 minutes to 151 minutes) over all visits.

Discussion: Simulated implementations of the predictive algorithm decreased length of stay but increased ordering costs. Optimal implementation of these predictions to reduce patient length of stay without incurring additional costs requires more exploration.

Conclusions: It is possible to predict common clinical orders placed during an ED visit with data available at triage.

Key words: machine learning, emergency medicine, clinical decision support systems

INTRODUCTION

Extended length of stay (LOS) in the emergency department (ED) is associated with adverse events including increased wait times, increased number of patients leaving before being seen, increased boarding rates, increased incidence of admission, increased risk of mortality, and reduced ability to respond to disasters.[1–4] Accordingly, EDs strive to maintain quality of care while reducing LOS. Data-driven clinical decision support tools have the potential to in-

crease efficiency of high-quality care in a fast-paced and resource constrained environment like the ED.[5,6] Researchers and clinicians have developed several ED focused tools, a subset of which use machine learning, to help with decision making and resource planning.[7] ED prediction models using machine learning include estimations of triage score,[8] clinical conditions such as urinary tract infections or sepsis,[9,10] patient LOS,[11] probability of patient admission from the ED,[12] number of admissions per hour,[13,14] and the risk of 72-hour ED revisits.[15] With appropriate implementation, ED clinical decision support tools could increase efficiency and reduce variability in decision making surrounding patient trajectories and resource allocation.[16] Furthermore, prospective models allow healthcare providers to allocate resources in a proactive, rather than reactive, manner before observing a measurable impact on patient flow.

Healthcare providers frequently place clinical orders after a patient has been assigned an ED bed, which may occur late in the patient visit. Often, a provider must wait for the results of orders before deciding on a treatment plan and disposition. Initiating orders as early as possible in the patient visit, such as at the time of triage, has the potential to shorten ED LOS by reducing wait time for clinical order results. Additionally, providing visit-specific recommended lists of orders may increase completeness and reduce variation in orders with respect to differential diagnoses.

Previous work has addressed the optimization and prediction of clinical orders. One intervention places a physician, rather than a lower-level provider, in triage to initiate clinical ordering.[17] Although this procedure ensures that orders are placed earlier in the patient visit, ED physicians are costly personnel who could be utilized elsewhere in the ED.

Order sets based on clinical guidelines are also a common tool for optimizing ordering.[18] These sets are frequently associated with a chief complaint. Once defined, order sets are straightforward to implement within modern electronic health record (EHR) systems. However, because order sets do not automatically update for changes in ordering guidelines or the availability of new orders, order sets require significant manual intervention to maintain. Guideline-based order sets are also not tailored to individual patients.

Finally, researchers have developed predictive models to anticipate ordering behavior for individual patient visits.[19,20] These methods provide an automated, data-driven, and personalized mechanism for predicting ordering behavior that have the potential to improve patient outcomes when compared with standard order sets.[21] However, previous methods dealt with non-ED clinical domains, focused on a specific subset of clinical conditions, or predicted all orders independently. Incorporating ED-specific context, encompassing a larger subset of clinical conditions, and incorporating knowledge of correlations between clinical orders could enhance the performance of predictive models describing ED clinical ordering behavior.[22]

We analyzed ED visits from 1 ED and 2 urgent care centers (UCCs) within the VA Boston Healthcare System over 56 months. For each visit, we extracted information available at triage including vital signs and chief complaint as well as previous medical history stored in the EHR. Using multilabel machine learning techniques, we trained site-specific models and predicted clinical orders for visits both independently and simultaneously and compared the algorithmic performance. We aim to provide predictive models that can be used to provide a patient-specific recommended list of orders to improve ordering ease and efficiency.

MATERIALS AND METHODS

Using only data available at the time of triage, we benchmarked 4 machine learning classifiers, including partial least squares classification, support vector machine, random forest, and multilayer perceptron, to predict clinical orders placed during an ED visit. We compared techniques that predict orders independently and simultaneously in a multilabel machine learning framework, attempting to harness the correlations between orders. We then quantified performance using classification performance metrics, including the F_1 score and the area under the receiver-operating characteristic curve (AUC). Additionally, we analyzed variable importance and simulated the effect of prediction on cost of ordering and LOS per visit. All portions of this study were approved by the institutional review board of the VA Boston Healthcare System (protocol #3059).

Study setting and population

We analyzed ordering behavior in 3 sites in the VA Boston Healthcare System, 1 ED and 2 UCCs, which we refer to as the ED, UCC 1, and UCC 2. Data were analyzed starting on April 12, 2012 (when the current instantiation of the ED Integration Software was installed), to December 31, 2016. The ED had a median visit volume of 13 345 visits per year with 13 beds and UCCs 1 and 2 had median yearly volumes of 10 770 and 5291 visits with 9 and 7 beds, respectively. Rates of admission were 35.06%, 26.39%, and 6.00% for the ED, UCC 1, and UCC 2, respectively. Visits were excluded if they were marked as entered in error or had a visit disposition indicating the patient had left before being seen, left against medical advice, eloped, or died during treatment. After excluding these visits, 140 855 (95.67%) visits remained across all 3 sites.

Data extraction and processing

We extracted information available to healthcare providers at the time of triage as well as clinical orders for ED and UCC visits from each site. Data was obtained from the Veterans Health Administration (VHA) Corporate Data Warehouse, a nationwide centralized data repository for researchers derived from data stored in the VHA's EHR.[23] Data elements available at triage consisted of the patient acuity, age, chief complaint, sex, order history for the patient's last visit at the site, problem list diagnoses, time of patient arrival, source of arrival, and vital signs.

All but 1 of the extracted data elements were structured. Structured data elements were either categorical or continuous variables. We extracted orders that occurred in at least 5% of visits and maintained a consistent label across the study period. The acuity assigned during triage, defined by the emergency severity index (ESI), was dichotomized into high (ESI 1, 2, and 3) and low (ESI 4 and 5) severity. The ESI algorithm is a reliable and validated triage tool used by the majority of EDs in the United States to determine patients' illness severity as well as the resources that will be needed to evaluate and treat the patient.[24,25] Patient age at the time of the ED visit was discretized into 5 bins following the National Hospital Ambulatory Medical Care Survey definitions.[26] Sex was binarized into female and male categories. For each order being predicted, we included a variable that indicated if that order had been placed during the patient's previous visit to the same ED or UCC, if a previous visit had occurred. Problem list diagnoses were consolidated into major diagnostic categories (MDCs), and we included a binary variable for each MDC indicating whether a diagnosis within that MDC was present on the patient's problem list at the time of the visit.[27] We also included 3 binary variables indicating whether the patient

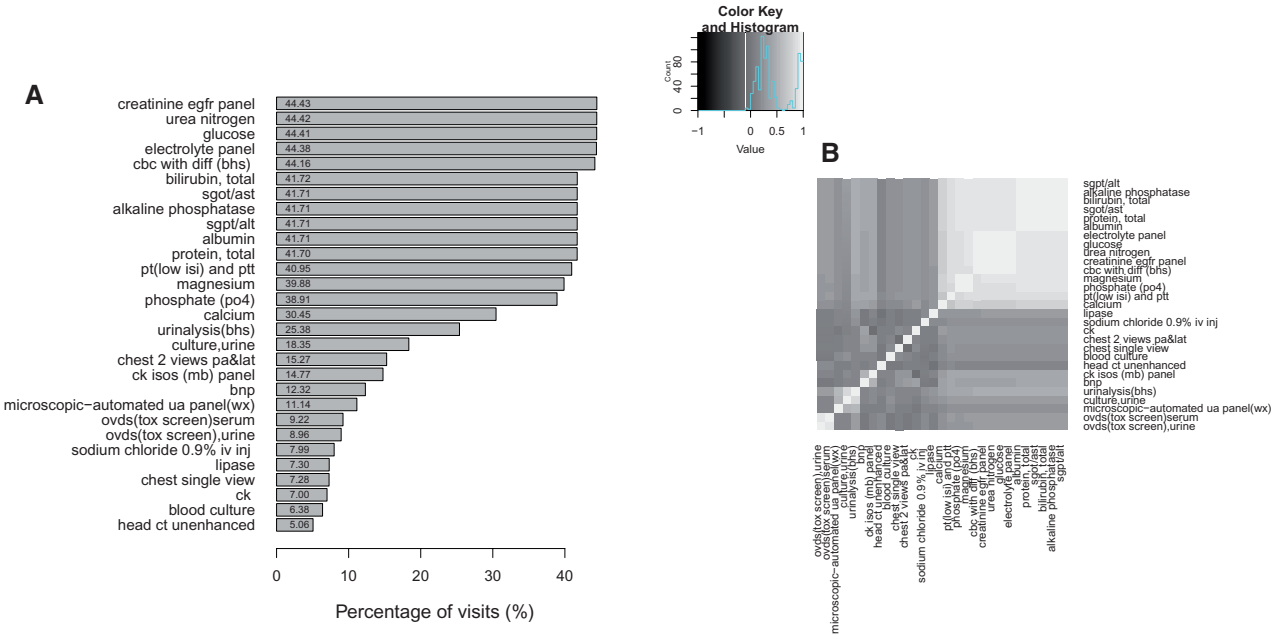

Figure 1. (A) Percentage of visits in which an order was placed from the emergency department (ED) site over the entire cohort of 64 427 visits. (B) Pearson's correlation coefficient between each pair of clinical orders across all visits to the ED.

arrived during the day (8:00 AM to 3:59 PM), evening (4:00 PM to 11:59 PM), or night (12:00 AM to 7:59 AM) and a variable for each unique source of patient arrival. Vital signs (blood pressure, pulse, pulse oximetry, respiration, temperature) and pain were discretized into very low, mildly low, normal, mildly high, or very high ranges, as applicable, and a binary variable was created for each category and vital sign pair (Supplementary Table S1).

Chief complaints were the only included data stored as unstructured free text. To incorporate chief complaints as variables in the model, we developed an automated natural language processing pipeline to map chief complaint text to standardized concepts in the Unified Medical Language System (UMLS).[28] We first preprocessed free text chief complaints by converting to lower case, removing punctuation, and trimming excess spaces. We then extracted a short form to long form mapping of emergency medicine specific abbreviations from the article abstracts of *Academic Emergency Medicine*, *Advanced Emergency Nursing Journal*, *Annals of Emergency Medicine*, *International Journal of Nursing Knowledge*, and *Journal of Emergency Nursing* from January 1, 2010, to June 30, 2018, using a previously published algorithm.[29] We replaced abbreviation short forms within chief complaints, using the frequency of the long form in the chief complaint corpus to disambiguate ambiguous abbreviation short forms (eg, "CP" was replaced with "chest pain" rather than "calprotectin"). As a final preprocessing step, we ran a spell check program, Hunspell.[30] We augmented the default English language dictionary with clinical terms from the UMLS and reorganized suggestions for misspellings by frequency of terms among all chief complaints (eg, "abess" was replaced by "abscess" rather than "abbess").[31] After the previous preprocessing steps were completed, we coded each chief complaint by performing partial string matching between chief complaint text and the Consumer Health Vocabulary and the DXplain vocabulary terms. Each chief complaint was then associated with zero or more UMLS concepts matching any single or multiword expression in the chief complaint. For example, the chief complaint "sore throat/cough" was associated with the concept identifiers C0234233 ("sore to touch"), C3665375 ("throat"), C0242429 ("sore throat"), and C0010200 ("coughing").

To improve predictive performance and increase training speed, we removed variables with near zero variance, variables that were highly correlated with other variables, and variables that formed linear dependencies with other variables in the dataset. Missing data, which constituted <5% of all data elements at each site, were replaced by the median value of that input variable.

Machine learning frameworks

Multilabel machine learning frameworks are applicable to problems that require each sample in a dataset to be assigned a subset of predefined labels.[32] For example, in our application, each ED visit is represented by a vector comprised of visit characteristics available at the time of triage, which must be mapped to a subset of clinical orders to be placed during the visit. We compared 2 multilabel machine learning frameworks for predicting clinical orders. We first predicted all orders independently with single-label binary classification techniques, a multilabel framework known as binary relevance. This formulation does not incorporate information contained in the correlations between labels. The second framework, known as random *k*-labelsets (RA*k*EL), employs an ensemble of multiclass classifiers.[22] To construct each multiclass classifier, the method selects a subset of all labels and defines a class for each unique combination of labels present in the training data. After training a predefined number of multiclass classifiers, each with a randomly drawn set of labels, a final decision for each label is made by voting over the output of all multiclass classifiers containing the label. Because RA*k*EL predicts subsets of labels simultaneously, the framework provides a mechanism for incorporating information from correlations between different labels.

Both binary relevance and RA*k*EL multilabel frameworks require the use of underlying binary or multiclass classifiers.

Table 1. Visits associated with each variable value used in the clinical order prediction model

Category	Variable	ED		UCC 1		UCC 2	
		n	%	n	%	n	%
Acuity	High	37 341	58.53	18 621	36.28	5794	22.69
	Low	26 457	41.47	32 702	63.72	19 741	77.31
Age	16-24 y	740	1.16	7	0.01	5	0.02
	25-44 y	9492	14.85	9028	17.56	4149	16.25
	45-64 y	20 571	32.19	15 119	29.41	6897	27.01
	65-74 y	16 391	25.65	14 492	28.19	7088	27.76
	75+ y	16 712	26.15	12 768	24.83	7396	28.96
Complaint concept (top 5)	Pain	18 010	28.00	9433	17.89	7192	27.15
	Chest	3606	5.61	1627	3.08	749	2.83
	Shortness of breath	3400	5.29	1009	1.91	431	1.63
	Abdomen	3243	5.04	677	1.28	447	1.69
	Chest pain	2841	4.42	761	1.44	507	1.91
Sex	Male	59 259	92.73	47 363	92.12	23 193	90.83
	Female	4647	7.27	4051	7.88	2342	9.17
Order history (top 5)	Creatinine eGFR panel	18 902	29.58	9569	18.61	2093	8.20
	Glucose	18 898	29.57	9549	18.57	2086	8.17
	Urea nitrogen	18 898	29.57	9567	18.61	2088	8.18
	Electrolyte panel	18 881	29.54	9611	18.69	2090	8.18
	CBC with diff (BHS)	18 796	29.41	9524	18.52	1607	6.29
Problem list (top 5)	Health status factors	42 639	72.19	36 830	76.29	17 892	75.32
	Musculoskeletal and connective tissue	40 431	68.45	31 428	65.10	16 283	68.55
	Endocrine, nutrition, metabolic	39 877	67.52	31179	64.58	14 947	62.92
	Circulatory system	39 526	66.92	29 106	60.29	13 708	57.71
	Mental diseases and disorders	34 813	58.94	33 684	69.77	16 119	67.86
Shift	Day	37 850	59.23	35 427	68.91	24 931	97.63
	Evening	19 786	30.96	11 070	21.53	184	0.72
	Night	6270	9.81	4917	9.56	420	1.64
Source of arrival (top 5)	Nonreferred	17 421	82.30	23 097	80.64	6179	87.97
	Offsite clinic	1442	6.81	0	0.00	56	0.80
	Other	1112	5.25	0	0.00	0	0.00
	Onsite clinic	463	2.19	769	2.68	506	7.20
	Offsite nursing home	268	1.27	0	0.00	0	0.00
Vital signs	Pulse oximetry normal	57 687	98.64	46 392	99.19	22 765	99.31
	Systolic normal	48 294	81.80	40 820	85.01	19 388	83.96
	Pulse normal	47 193	79.85	38 906	81.01	18 834	81.36
	Diastolic normal	46 638	79.03	37 730	78.60	18 772	81.36
	Respiration normal	45 737	77.50	44 008	91.79	20 140	87.34
	No pain	24 714	42.42	23 232	48.85	7462	32.67
	Temperature normal	17 632	30.08	11 905	25.06	7561	33.05

BHS: Boston Healthcare System; CBC: complete blood count; ED: emergency department; eGFR: estimated glomerular filtration rate; UCC: urgent care center.

We benchmarked 4 classifiers with each framework: partial least squares classification, support vector machine, random forest, and MLP.

Model performance evaluation

All analyses were conducted with the R statistical programming language version 3.5.1 using the VA Informatics and Computing Infrastructure.[33] We wrote a custom R implementation of the RA*k*EL framework and used the R package *caret* to run the 4 classification methods.[34] Hyperparameters for each classifier, if applicable, and classification thresholds for output were selected by performing a grid search with 10-fold cross-validation and selecting hyperparameter values that maximized Cohen's kappa statistic.[35]

For all classifiers, we trained on a randomly sampled 50% of 2 contiguous months of data, weighting the sampling of training visits across the 2 months by the inverse of the number of orders per visit to oversample for lower frequency orders. We then tested the trained model on all visits in the subsequent month to mimic an online learning paradigm that would occur in the ED environment. This data split resulted in 18 unique training-testing replicate pairs. Additionally, for RA*k*EL, we selected subsets of orders in each multiclass classifier with weighted sampling corresponding to the inverse frequency of the order to similarly oversample lower frequency orders in the training procedure.

To evaluate predictive performance, we utilized the AUC and the F_1 score, the harmonic mean between precision and recall. We calculated the F_1 score for each order individually as well as the median across all orders. AUC was calculated as the median value across all orders.

In addition to calculating metrics of predictive performance, we also calculated relative variable importance measures for individual orders using *caret*'s "filter" approach.[34] This method determines the AUC of using each variable alone to predict the outcome variable and then scales the AUC values across all variables from 0 to 100.

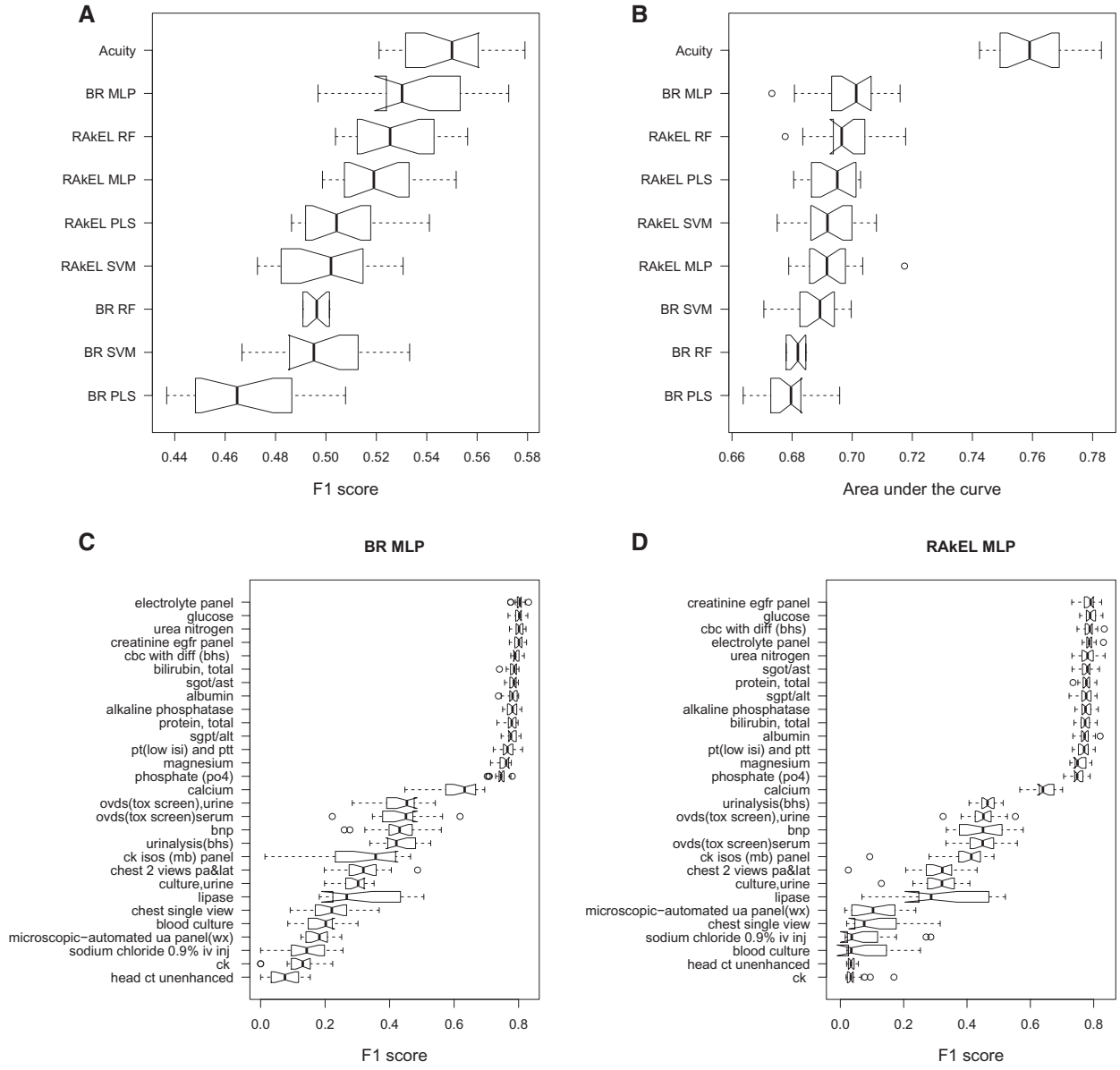

Figure 2. Distribution of the (A) F_1 score and (B) area under the receiver-operating characteristic curve for all methods over all orders. F_1 score for the prediction of each order individually with (C) binary relevance framework and the (D) random *k*-labelsets (RA*k*EL) framework using a multilayer perceptron. BR: binary relevance; MLP: multilayer perceptron; PLS: partial least squares; RF: random forest; SVM: support vector machine.

Raw predictive performance is not the only element to consider when utilizing predictive algorithms to suggest clinical orders. We also sought to investigate whether the costs of additional orders caused by false positive predictions do not overwhelm the benefits of reduced LOS, when all orders are successfully predicted for a visit. To explore this issue, we constructed a deterministic simulation of the effect of order prediction in terms of order cost per visit and visit LOS. The inputs for this simulation are the list of visits in the test set of each replicate for the ED site, the LOS for those visits, the predicted orders for those visits, and the orders for that visit that occurred in reality.

To determine the cost of using order predictions we assess which tests are predicted for a given patient. The costs of those tests are calculated and compared with the total cost of the tests that were ordered for the patient in reality. Assuming that the provider had indeed ordered the correct tests, the difference between these costs summed across all patients represents the total increased cost incurred by automating the process using our prediction models. Lab order costs were extracted from values recorded with respect to specific orders in the EHR. Radiology order costs were assessed by taking the assigned relative value units for the radiology order and calculating the cost of work by multiplying the relative value unit by the conversion factor and geographic practice cost index from the metropolitan Boston area for 2017.[36]

To analyze the potential trade-off of any increase in ordering costs with decreased LOS, we adjusted LOS according to simplified model

Table 2. Performance metrics for evaluating each framework and classifier model pair in addition to using acuity alone to predict orders

	F_1	Recall	Precision	False positive rate	Accuracy	AUC
BR PLS	0.46	0.51	0.62	0.16[a]	0.84[a]	0.68
BR MLP	0.53	0.57	0.52	0.17	0.83	0.70
BR SVM	0.50	0.53	0.57	0.16[a]	0.84[a]	0.69
BR RF	0.50	0.51	0.71[a]	0.16[a]	0.84[a]	0.68
RA*k*EL PLS	0.50	0.57	0.59	0.18	0.84[a]	0.70
RA*k*EL MLP	0.52	0.60	0.54	0.21	0.82	0.69
RA*k*EL SVM	0.50	0.57	0.60	0.18	0.84[a]	0.69
RA*k*EL RF	0.53	0.60	0.52	0.20	0.83	0.70
Acuity	0.55[a]	0.95[a]	0.43	0.44	0.66	0.76[a]

AUC: area under the receiver-operating characteristic curve; BR: binary relevance; MLP: multilayer perceptron; PLS: partial least squares; RA*k*EL: random *k*-labelsets; RF: random forest; SVM: support vector machine.

[a]Best value for performance metric.

of ED visits. ED visits were modeled as progressing through 5 time points: arrival, triage, first seen by a provider, disposition set, and departure. Timestamps marking these events are used in national VHA quality metrics for EDs and UCCs and are reliably recorded from EHR activity. We simulated the potential effects of predicting orders on LOS by modifying the real visit LOS according to the accuracy of the predictive algorithm (Supplementary Figure S1). The historic visits were divided into 3 separate groups (A, B, and C). Group A contains patients whose full list of orders was not predicted by the algorithm. The LOS for group A patients remained the same as the observed LOS for their visits. Groups B and C contain patients for which all orders were predicted by the algorithm. Group B patients are those whose runtime of placed orders was less than or equal to the time between triage and the time the patient was first seen. For group B patients, we adjusted the time between when the patient was first seen and the disposition time while retaining the actual recorded time lapse for all other intervals. The time between seen and disposition was sampled from an exponential distribution with a mean of 51 minutes,

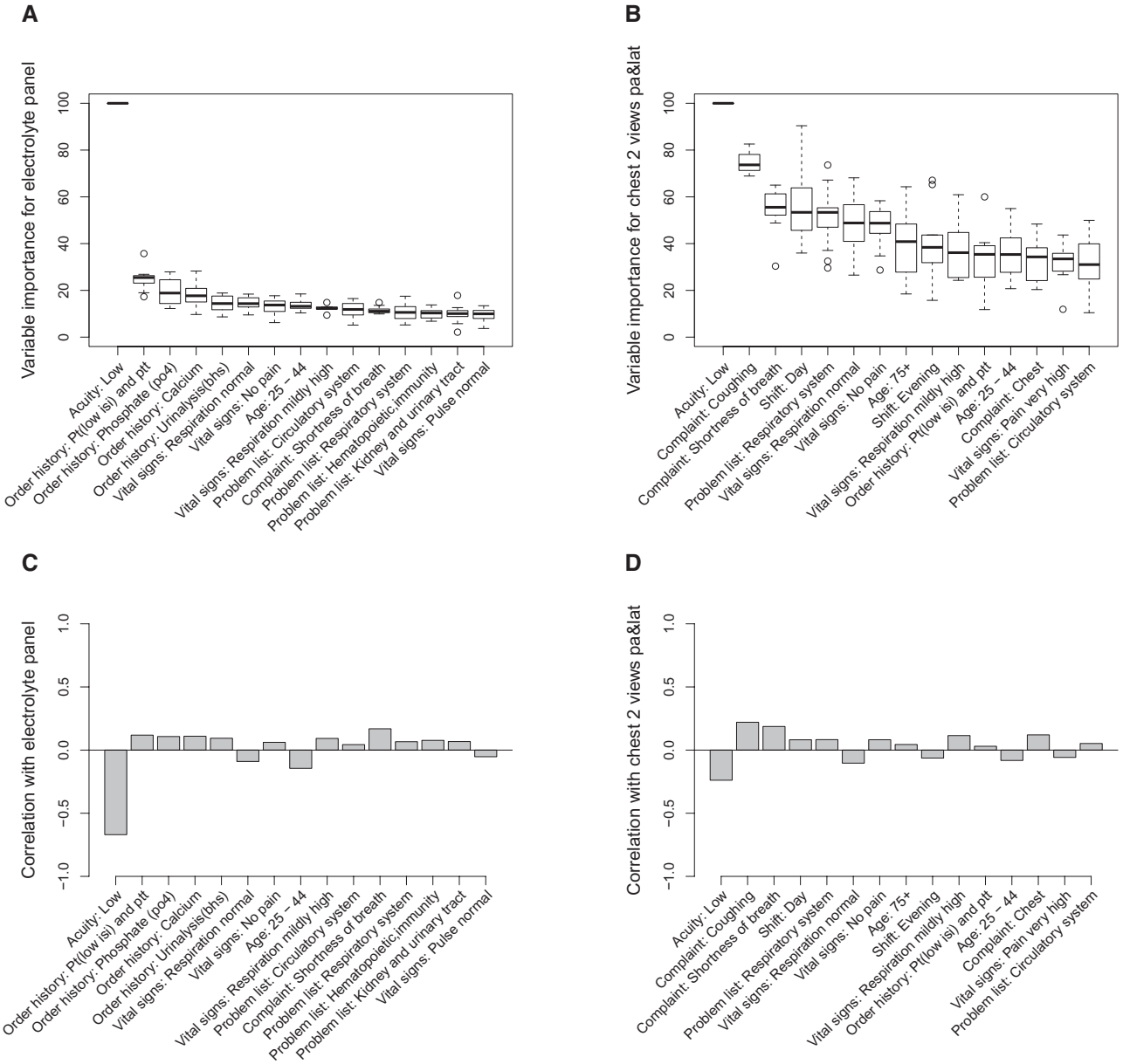

Figure 3. Scaled variable importance for prediction for (A) an electrolyte panel and (B) a chest x-ray order. Pearson's correlation coefficient for the top 15 most important input variables with (C) an electrolyte panel and (D) a chest x-ray order.

a distribution that approximately matched seen to disposition time for patients with no orders in the dataset (Supplementary Figure S2). Group C patients are those whose runtime of placed orders was greater than the time between triage and the time the patient was first seen. For group C patients, we extended the time between triage and time first seen to the maximum order runtime and then adjusted the time between when the patient was first seen and the disposition time as described for group B. The distributions of observed and adjusted LOS were then compared. Because of random sampling, we computed 95% confidence intervals over 10 runs of the simulation to ensure stability of the median simulated LOS estimate.

Last, we generated examples of order sets by chief complaint by assessing the percentage of visits with a predicted order within a certain chief complaint group.

RESULTS

Order prevalence and correlation structure for the ED are shown in Figure 1. Orders varied both in prevalence and in their correlation to other orders, with strong correlation structures indicating bundled orders. For the ED site over the nearly 5 years of the dataset, orders were placed for 2179 unique orderable items. Of these items, 594 (27%) were ordered only once and 1378 (63%) were ordered <10 times. There were 29 orders for the ED site that were consistently labeled and occurred in at least 5% of visits. We used these 29 orders as prediction targets. We note that these 29 orders represented 81% of all order instances in the ED over the entire cohort time as most unique orderable items are rarely ordered (Supplementary Figure S3). Summary statistics for each analyzed site are shown in Table 1. Sites varied in terms of patient volume, condition severity, and percentage of visits admitted but have comparable patient demographics with respect to age and sex. As expected, the ED saw a greater percentage of high acuity patients than either of the UCCs.

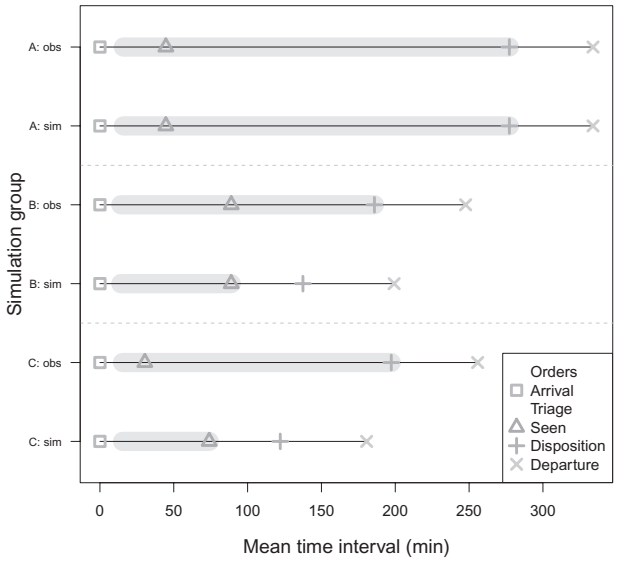

Figure 4. Summary of simulation procedure results for assessing the effect of the prediction algorithm on reducing length of stay on each simulation group. Mean time interval estimates are shown for each group along with a time range in which orders are placed and run (gray shading) for both the original observed (obs) visit time intervals and the simulated (sim) visit time intervals. Group definitions are as follows: (A) if all orders are not correctly predicted, length of stay remains the same; (B) when all orders are correctly predicted at triage and the time between triage and time seen exceeds the amount of time for maximum order runtime, the time between when the patient is first seen and when the disposition is sampled from an exponential distribution; and (C) when all orders are correctly predicted at triage and the time between triage and time seen is less than the maximum order runtime, we extend the time between triage and first seen and sample the time between when the patient is first seen and the disposition from an exponential distribution.

Overall, 95.83% of free text chief complaints across the cohort were associated with at least 1 UMLS concept. The pipeline associated an average of 3.03 clinical concepts per chief complaint among complaints with at least 1 associated concept. The concept C0030193 ("pain") was the most commonly associated clinical concept, identified in 28.00% of chief complaints. The prevalence of the top 20 clinical concepts occurring in chief complaints is shown in Supplementary Figure S4. Examples of raw and standardized chief complaints are shown in Supplementary Table S2.

Comparison of both the AUC and the F_1 score metrics reveals that multilabel models using the RA*k*EL multilabel framework frequently outperform the binary relevance framework (Figure 2A, B). However, the binary relevance framework using the multilayer perceptron outperformed all other methods (F_1 score = 0.53, AUC = 0.70). The binary relevance framework using the partial least squares classifier demonstrated the least predictive power (F_1 score = 0.46, AUC = 0.68). Looking at individual order performance revealed that the RA*k*EL framework slightly outperformed binary relevance for highly correlated orders (Figure 2C, D). For example, on the 14 most prevalent orders, the binary relevance framework using the multilayer perceptron achieved a median F_1 score of 0.78 while the RA*k*EL framework utilizing a random forest achieved 0.79. In general, predictive performance for individual orders is highly correlated with the frequency of the order in the dataset (Supplementary Figure S5). Comparisons of order frequency and the number of orders per visit between observed and predicted data are shown in Supplementary Figures S6 and S7. In general, models tend to underorder for rarer orders and overorder for common orders.

Additional prediction metrics (eg, recall and precision) are shown in Table 2. We also compare to using the acuity score alone as a classifier. If the visit has high acuity, all tests are ordered for a visit. If the visit has low acuity, no tests are ordered. Using acuity alone as a proxy for ordering results in substantial overordering of tests. For example, the false positive rate increases from 17% for binary relevance (BR) multilayer perceptron (MLP) to 44% when using acuity to predict orders. Additionally, we show precision-recall curves for 2 orders using the BR MLP model in Supplementary Figure S8. Although performance is noisy for low performing order tasks, precision-recall curves for well predicted orders demonstrate consistent performance with a sharp drop of recall at a given threshold, indicating that there is a narrow window for thresholding model output for calling order predictions.

Variable importance for multilabel models varied by clinical order, but visit acuity was highly predictive. Orders placed on previous ED visits were important for predicting bundled clinical orders such as the electrolyte panel (Figure 3A, C). Radiology orders (eg, "chest 2 views pa&lat") were additionally correlated with the time of day in which the patient visits the ED (Figure 3B, D).

Over 20 264 visits in the ED testing sets, the BR MLP correctly predicted all orders placed before each visit's disposition decision and had at least 1 order placed in 2030 (10.02%) visits. Figure 4 shows the mean for each time interval across visits grouped as described in the methods section. According to our simulation, median costs over all visits increased from $21 to $45 per visit, but with overlapping interquartile ranges, while median visit LOS decreased from 158 minutes to 151 minutes (95% confidence interval, 151.01-151.66 minutes) (Figure 5). A 2-sided Wilcoxon test indicated that

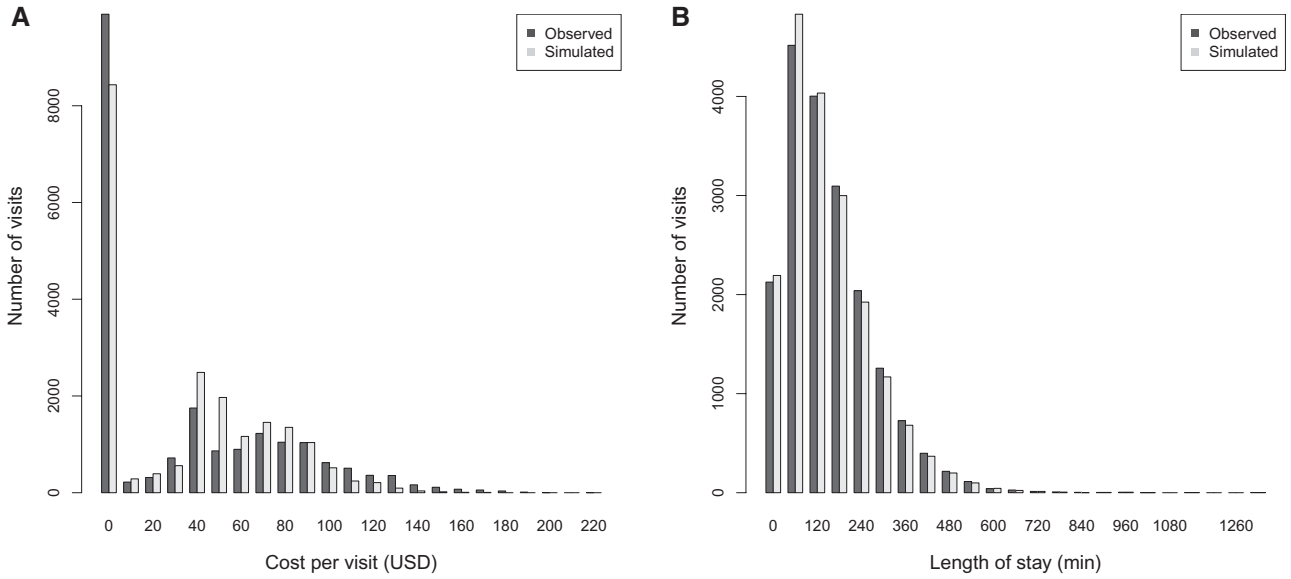

Figure 5. Simulation of the effects of the predictive algorithm on visits to the emergency department site on all visits the test sets. Simulation effects are measured in terms of (A) total cost of orders placed during a visit, binned in $10 increments, and (B) length of stay per visit, binned in 60-minute increments. USD: U.S. dollars.

distributions of costs ($P = 1.46 \times 10^{-4}$) and LOS ($P = 1.36 \times 10^{-58}$) per visit are significantly different between observed and simulated visits. Based on the median surveyed VHA ED physician wage in 2015 ($208 891 annual or approximately $100 an hour), the median additional cost from prediction ($24) would be equivalent to 14.4 minutes of a physician's time.[37] Therefore, if time between time seen and disposition could be shortened by at least 15 minutes, the extra ordering cost would be compensated by saved provider time. We also used acuity to determine ordering in the simulation framework. Using acuity, 43% of visit were eligible for LOS modification. Across all visits, median LOS decreased from 158 minutes in the observed data to 133 minutes after the simulation while median ordering cost increased from $21 to $229 per visit.

Consistency of prediction performance varied by the patient volume of each site across 56 months (Supplementary Figure S9). UCC 2, the site with the lowest patient volume, showed the greatest variance in F_1 score across time, ranging from 0.24 to 0.56. The ED, the site with the highest patient volume, showed smaller variance in F_1 score across time, ranging from 0.51 to 0.57.

Finally, we contrasted predicted order sets by standardized chief complaint terms, as shown in Figure 6, to generate recommendations for standard order sets. Commonly predicted orders vary by chief complaint. A chest x-ray ("chest 2 views pa&lat") is more often predicted for visits with "shortness of breath" than for "chest pain." Toxicology screening is only predicted in over 50% of orders for visits with chief complaints containing the standardized term "ethanol." Visits involving "refills" generally have no predicted orders.

DISCUSSION

There are several potential applications of the proposed model. Predictive algorithms such as the one described could be used to design data-driven ordering sets. Additionally, the algorithm can highlight which orders are highly predictable for a given subset of patients and potentially suitable for standing orders. The method could also be used to recommend visit specific lists of orders for a given patient, facilitating the order selection process for clinicians.

Although using acuity alone as a proxy for ordering shows the best AUC and F_1 scores, further analysis with other performance metrics and the simulation indicate that using acuity alone would increase the false positive rate to an unreasonable level. Given that excessive ordering of blood tests and imaging is associated with adverse patient outcomes such as hospital-acquired anemia, exposure to radiation, and burden of follow-up on false positive results, ordering purely on the basis of acuity has a substantial disadvantage.[38,39] In addition, simulation results indicate that using acuity alone would increase ordering costs by a factor of 10 ($21-$229 per visit). Although using multilabel machine learning models also result in false positives, higher cost orders are often underordered relative to low-cost orders. The multilabel models provide a more reasonable compromise between LOS and cost than using acuity alone.

The study has several limitations. There is an opportunity to increase the scope of our study to more sites to better explore the generalizability of our conclusions. Our study only analyzed 3 sites of over 130 EDs and UCCs across the country run by the VA and over a period of <5 years. Similarly, it is worth exploring how our approach would apply to private health systems, which may have different ordering behavior based on different payment structures. Another limitation of our work is that we only analyzed relatively common orders that were systematically named. To address the issue of heterogeneous data collection, the VA has several initiatives to consolidate order names, along with other data elements, including transitioning records from administrative relational databases into the Observational Medical Outcomes Partnership common data model.[40] Last, the current predictive framework requires retraining at each site or time frame with different combinations of clinical orders. However, the models can be retrained on new datasets in an automated and efficient fashion.

CONCLUSION

We applied multilabel machine learning algorithms that predict ordering behavior in an ED visit with data available at the time of

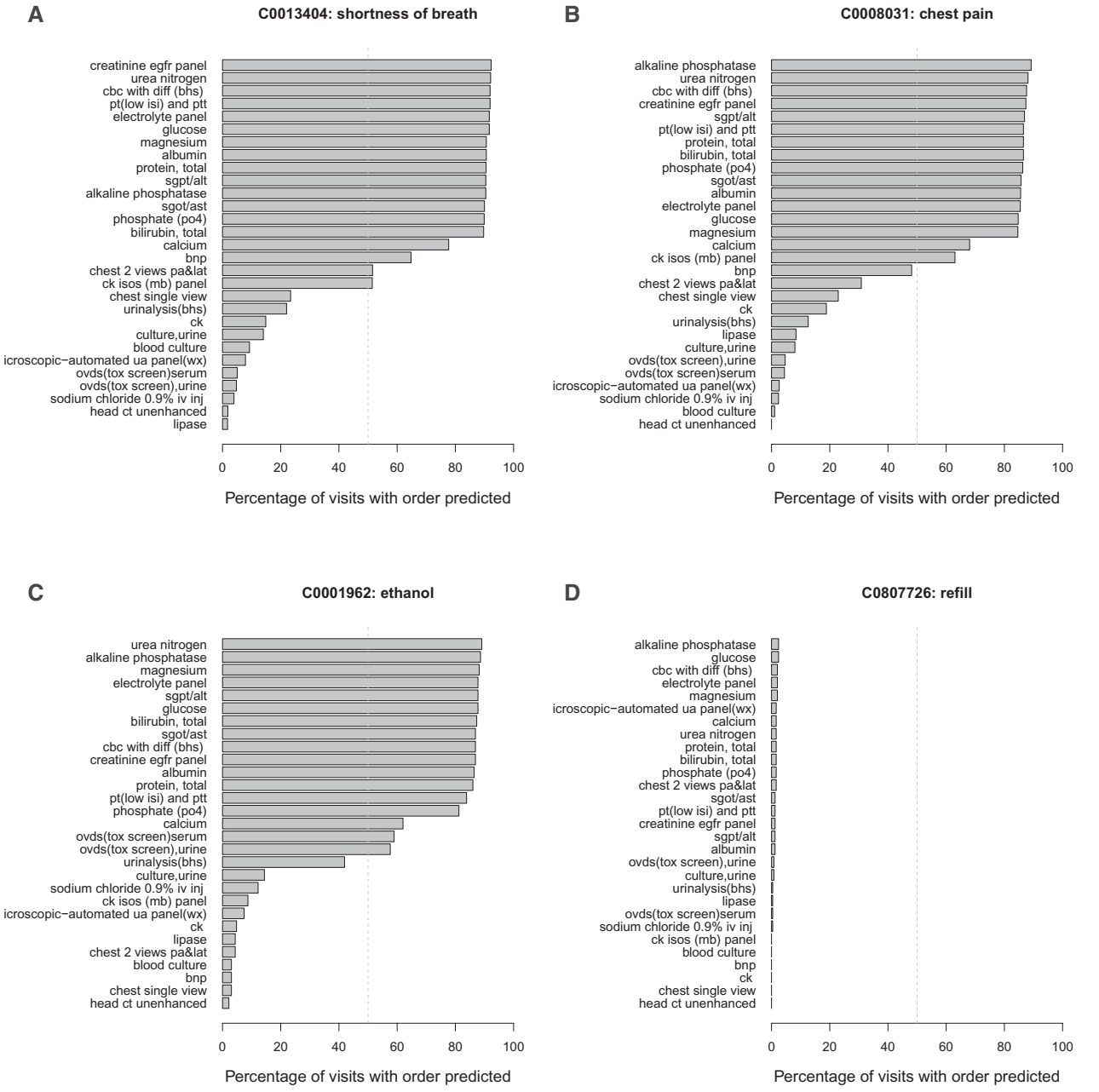

Figure 6

Figure 6. Percentage of visits with a predicted order for visits with chief complaints with standardized terms (A) C0013404: "shortness of breath," (B) C0008031: "chest pain," (C) C0001962: "ethanol," and (D) C0807726: "refill," Only visits from the ED site in the test sets were used. The dashed gray line indicates 50%.

triage. Although predicting orders independently produced superior results for rarer orders, using machine learning models that simultaneously predict clinical orders for a given visit enhanced predictive performance for common orders, indicating that accounting for correlations more realistically models clinical ordering behavior. Overall, predictive performance is higher for more frequently placed orders but is relatively stable across time and multiple institutions. According to a simulation of implementation, utilizing predictions would decrease visit LOS but increases costs of ordering per visit, indicating potential for algorithmic improvement.

FUNDING

This work was supported by the HHZ received the VA Special Fellowship in Medical Informatics from the U.S. Department of Veterans Affairs, resources and the use of facilities at the VA Boston

Healthcare System, and funds from the Maine Medical Center. The contents do not represent the views of the U.S. Department of Veterans Affairs or the U.S. Government.

AUTHOR CONTRIBUTIONS

HSH-Z conducted data extraction and analysis of the data. HSH-Z and SAG drafted the manuscript. SAG and JSP acquired funding. All authors contributed to study concept and design, interpretation of data, and critical revision of the manuscript.

SUPPLEMENTARY MATERIAL

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

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Trzeciak S, Rivers EP. Emergency department overcrowding in the United States. *Emerg Med J* 2003; 20 (5): 402–5.
2. Gorski JK, Batt RJ, Otles E, *et al*. The impact of emergency department census on the decision to admit. *Acad Emerg Med* 2017; 24 (1): 13–21.
3. Villa-Roel C, Guo X, Holroyd BR, *et al*. The role of full capacity protocols on mitigating overcrowding in EDs. *Am J Emerg Med* 2012; 30 (3): 412–20.
4. Singer AJ, Thode HC, Viccellio P, *et al*. The association between length of emergency department boarding and mortality. *Acad Emerg Med* 2011; 18 (12): 1324–9.
5. Puskarich MA, Callaway C, Silbergleit R, *et al*. Priorities to overcome barriers impacting data science application in emergency care research. *Acad Emerg Med* 2018; 26(1): 97–105.
6. Stewart J, Sprivulis P, Dwivedi G. Artificial intelligence and machine learning in emergency medicine. *Emerg Med Australas* 2018; 30 (6): 870–4.
7. Bennett P, Hardiker NR. The use of computerized clinical decision support systems in emergency care: a substantive review of the literature. *J Am Med Inform Assoc* 2017; 24: 655–68.
8. Levin S, Toerper M, Hamrock E, *et al*. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. *Ann Emerg Med* 2018; 71 (5): 565–74.e2.
9. Taylor RA, Moore CL, Cheung KH, *et al*. Predicting urinary tract infections in the emergency department with machine learning. *PLoS One* 2018; 13 (3): e0194085.
10. Horng S, Sontag DA, Halpern Y, *et al*. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS One* 2017; 12 (4): e0174708.
11. Gill SD, Lane SE, Sheridan M, *et al*. Why do 'fast track' patients stay more than four hours in the emergency department? An investigation of factors that predict length of stay. *Emerg Med Australas* 2018; 30 (5): 641–7.
12. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. *PLoS One* 2018; 13 (7): e0201016.
13. Peck JS, Benneyan JC, Nightingale DJ, *et al*. Predicting emergency department inpatient admissions to improve same-day patient flow. *Acad Emerg Med* 2012; 19: 1045–54.
14. Peck JS, Gaehde SA, Nightingale DJ, *et al*. Generalizability of a simple approach for predicting hospital admission from an emergency department. *Acad Emerg Med* 2013; 20 (11): 1156–63.
15. Pellerin G, Gao K, Kaminsky L. Predicting 72-hour emergency department revisits. *Am J Emerg Med* 2018; 36 (3): 420–4.
16. Peck JS, Benneyan JC, Nightingale DJ, *et al*. Characterizing the value of predictive analytics in facilitating hospital patient flow. *IIE Trans Healthc Syst Eng* 2014; 4 (3): 135–43.
17. Rowe BH, Guo X, Villa-Roel C, *et al*. The role of triage liaison physicians on mitigating overcrowding in emergency departments: a systematic review. *Acad Emerg Med* 2011; 18 (2): 111–20.
18. Zhang Y, Padman R, Levin JE. Paving the COWpath: data-driven design of pediatric order sets. *J Am Med Inform Assoc* 2014; 21 (e2): e304–11.
19. Chen JH, Podchiyska T, Altman RB. OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records. *J Am Med Inform Assoc* 2016; 23 (2): 339–48.
20. Klann JG, Szolovits P, Downs SM, *et al*. Decision support from local data: creating adaptive order menus from past clinician behavior. *J Biomed Inform* 2014; 48: 84–93.
21. Wang JK, Hom J, Balasubramanian S, *et al*. An evaluation of clinical order patterns machine-learned from clinician cohorts stratified by patient mortality outcomes. *J Biomed Inform* 2018; 86: 109–19.
22. Tsoumakas G, Katakis I, Vlahavas I. Random k-Labelsets for multi-label classification. *IEEE Trans Knowl Data Eng* 2011; 23 (7): 1079–89.
23. Fihn SD, Francis J, Clancy C, *et al*. Insights from advanced analytics at the veterans health administration. *Health Aff (Millwood)* 2014; 33 (7): 1203–11.
24. Gilboy N, Tanabe T, Travers DR, Gilboy N, Travers D, Rosenau AT. *Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, Version 4. Implementation Handbook*. Rockville, MD: *Agency for Healthcare Research and Quality*; 2011.
25. Tanabe P, Gimbel R, Yarnold PR, *et al*. Reliability and validity of scores on the emergency severity index version 3. *Acad Emerg Med* 2004; 11 (1): 59–65.
26. McCaig LF, Burt CW. Understanding and interpreting the national hospital ambulatory medical care survey: key questions and answers. *Ann Emerg Med* 2012; 60 (6): 716–21.e1.
27. Centers for Medicare and Medicaid Services. ICD-10-CM/PCS MS-DRG v36.0 Definitions Manual. https://www.cms.gov/ICD10Manual/version36-fullcode-cms/fullcode_cms/P0001.html Accessed August 22, 2019.
28. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32 (database issue): D267–70.
29. Yu H, Hripcsak G, Friedman C. Mapping abbreviations to full forms in biomedical articles. *J Am Med Inform Assoc* 2002; 9 (3): 262–72.
30. Hornik K, Murdoch D. Watch your spelling! *R J* 2011; 3 (2): 22–8.
31. Crowell J, Zeng Q, Ngo L, Lacroix E-M. A frequency-based technique to improve the spelling suggestion rank in medical queries. *J Am Med Inform Assoc* 2004; 11 (3): 179–85.
32. Zhang ML, Zhou ZH. A review on multi-label learning algorithms. *IEEE Trans Knowl Data Eng* 2014; 26 (8): 1819–37.
33. VA Informatics and Computing Infrastructure (VINCI). VA HSR HIR 08-204, U.S. Department of Veterans Affairs. https://vaww.vinci.med.va.gov Accessed August 22, 2019.
34. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw* 2008; 28 (5): 1–26.
35. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas* 1960; 20 (1): 37–46.
36. U.S. Centers for Medicare and Medicaid Services. National Physician Fee Schedule and Relative Value Files. https://www.cms.gov/apps/physician-fee-schedule/documentation.aspx Accessed August 22, 2019.
37. Ward MJ, Collins SP, Pines JM, Dill C, Tyndall G, Kessler CS. Emergency medicine in the Veterans Health Administration-results from a nationwide survey. *Am J Emerg Med* 2015; 33 (7): 899–903.
38. Eaton KP, Levy K, Soong C, *et al*. Evidence-based guidelines to eliminate repetitive laboratory testing. *JAMA Intern Med* 2017; 177 (12): 1833–9.
39. Chung JH, Duszak R, Hemingway J, *et al*. Increasing utilization of chest imaging in US emergency departments from 1994 to 2015. *J Am Coll Radiol* 2019; 16 (5): 674–82.
40. Overhage JM, Ryan PB, Reich CG, *et al*. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc* 2012; 19 (1): 54–60.