OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records

RECEIVED 25 February 2015
REVISED 29 May 2015
ACCEPTED 3 June 2015
PUBLISHED ONLINE FIRST 21 July 2015

Jonathan H Chen[1,2], Tanya Podchiyska[3], Russ B Altman[4]

ABSTRACT

Objective: To answer a "grand challenge" in clinical decision support, the authors produced a recommender system that automatically data-mines inpatient decision support from electronic medical records (EMR), analogous to Netflix or Amazon.com's product recommender.

Materials and Methods: EMR data were extracted from 1 year of hospitalizations (>18K patients with >5.4M structured items including clinical orders, lab results, and diagnosis codes). Association statistics were counted for the ~1.5K most common items to drive an order recommender. The authors assessed the recommender's ability to predict hospital admission orders and outcomes based on initial encounter data from separate validation patients.

Results: Compared to a reference benchmark of using the overall most common orders, the recommender using temporal relationships improves precision at 10 recommendations from 33% to 38% ($P < 10^{-10}$) for hospital admission orders. Relative risk-based association methods improve inverse frequency weighted recall from 4% to 16% ($P < 10^{-16}$). The framework yields a prediction receiver operating characteristic area under curve (c-statistic) of 0.84 for 30 day mortality, 0.84 for 1 week need for ICU life support, 0.80 for 1 week hospital discharge, and 0.68 for 30-day readmission.

Discussion: Recommender results quantitatively improve on reference benchmarks and qualitatively appear clinically reasonable. The method assumes that aggregate decision making converges appropriately, but ongoing evaluation is necessary to discern *common* behaviors from "correct" ones.

Conclusions: Collaborative filtering recommender algorithms generate clinical decision support that is predictive of real practice patterns and clinical outcomes. Incorporating temporal relationships improves accuracy. Different evaluation metrics satisfy different goals (predicting likely events vs. "interesting" suggestions).

Keywords: clinical decision support systems, electronic health records, data-mining, recommender algorithms, collaborative filtering, practice guidelines, practice variability, order sets

BACKGROUND AND SIGNIFICANCE

Variability and uncertainty in medical practice compromise quality of care and cost efficiency. Knowledge is inconsistently applied, such as 25% of patients with a heart attack not receiving the aspirin they should and overall compliance with evidence-based guidelines ranging from 20% to 80%.[1] A majority of clinical decisions (e.g., a third of surgeries to place pacemakers or ear tubes) lack adequate evidence to support or refute their practice.[1] Even with disruptive reforms,[2] evidence-based medicine from randomized control trials cannot keep pace with the perpetually expanding breadth of clinical questions. Medical practice is thus routinely driven by individual expert opinion and anecdotal experience. The advent of the meaningful use era of electronic medical records (EMRs)[3] creates a new opportunity for *data-driven* clinical decision support (CDS) that utilizes the collective expertise of many practitioners in a learning health system.[4–8]

CDS tools such as order sets already help reinforce consistency and compliance with best practices,[9,10] but CDS production is limited in scale by a top-down, knowledge-based approach requiring the manual effort of human experts.[11] One of the "grand challenges" in CDS[12] is thus the automatic production of CDS from the bottom-up by data-mining clinical data sources. Instead of consulting individual experts for advice, a data-driven approach would allow us to effectively consult *every* practitioner, learning how they all care for their similar patients on a statistical average.

OBJECTIVE

Inspired by analogous information retrieval problems in recommender systems, collaborative filtering, market basket analysis, and natural language processing, we sought to automatically generate CDS content in the form of a clinical order recommender system[13,14] analogous to Netflix or Amazon.com's "customers who bought A also bought B" system.[15] While a clinician may consider a broad differential diagnosis of possibilities or the risks and benefits of many interventions, clinical orders (e.g., labs, imaging, medications) are the concrete manifestations of point-of-care decision making. Our broad vision is to seamlessly integrate a system into clinical order entry workflows that automatically infers the relevant clinical context based on data already in the EMR and provides actionable decision support in the form of clinical order suggestions. Prior approaches to automated CDS content development that predicts future or missing clinical orders includes association rules, nearest neighbors, logistic regression, Bayesian networks, and unsupervised clustering of clinical orders.[16–24] Our own preliminary work explored the impact of temporal relationships, alternative measures for uncommon but "interesting" recommendations, and generalization towards predicting clinical outcomes.[13,14] Here we provide a more extensive evaluation of these concepts to determine how robust the algorithm is for predicting more varied clinical outcomes, whether it can be improved with different approaches to item counts and score aggregation, and illustrate how

Correspondence to Russ B Altman, Shriram Room 209, MC: 4245, 443 Via Ortega Drive, Stanford, CA 94305-4145, USA; russ.altman@stanford.edu; Tel: 650-725-3394

Published by Oxford University Press on behalf of the American Medical Informatics Association 2015. This work is written by US Government employees and is in the public domain in the US. For numbered affiliations see end of article.

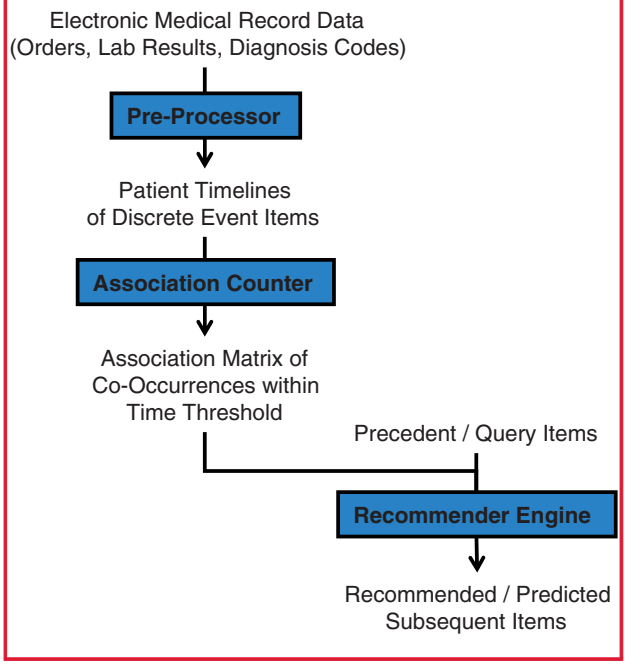
RESEARCH AND APPLICATIONS
Downloaded from http://jamia.oxfordjournals.org/ by guest on April 5, 2016

disease-specific inferences are possible even without specified diagnoses.

MATERIALS AND METHODS

Figure 1 outlines the data flow to implement and execute a clinical order recommender. To begin, de-identified, structured EMR data from all inpatient hospitalizations at Stanford University Hospital in 2011 was extracted by the STRIDE project.[25] These data cover patient encounters from their initial (emergency room) presentation until hospital discharge. With >18K patients, these data include >5.4M instances of >17K distinct clinical items, with patients, items, and instances, respectively, analogous to customers, items in a product catalog, and items actually purchased by a customer. The clinical item elements include >3500 medication, >1000 laboratory, >800 imaging, and >700 nursing orders. Nonorder items include >1000 lab results, >5800 problem list entries, >3400 admission diagnosis ICD9 codes, and patient demographics. Medication data were normalized with RxNorm mappings[26] down to active ingredients and routes of administration. The ICD9 coding hierarchy was rebuilt up to three digit codes, such that an item for code 786.05 would have additional items replicated for code 786.0 and 786. Numerical lab results were binned into categories based on "abnormal" flags established by the clinical laboratory. The above preprocessing models each patient as a timeline of clinical item instances as in Figure 2, with each instance mapping a clinical item to a patient at a discrete time point.

Figure 1: Algorithm flowchart. Structured data extracted from the electronic medical record (EMR) captures inpatient data from initial (emergency room) presentation to discharge. Data is preprocessed into timelines of clinical item instances, including normalization of medications to active ingredient, binning of numerical results into categories, and exclusion of rare and routine process orders. Association statistics are precomputed by counting item co-occurrences within a designated time threshold parameter. Given a set of query items, the recommender engine searches the association statistics to produce a score-ranked list of related orders.



With the clinical item instances following the "80/20 rule" of a power law distribution,[27] most clinical items may be ignored with minimal information loss. In this case, ignoring rare clinical items with <256 instances (0.005% of all instances) reduces the effective item count from >17K to ~1.5K (9%), while still capturing 5.1M (94%) of the 5.4M item instances. Process orders (e.g., vital signs, peripheral Intravenous (IV) care, patient weight, regular diet, transport patient, and all PRN medications) were excluded as they generally reflect routine care. The above leaves $m = 1482$ distinct clinical items, of which 808 are clinical orders. This reduction greatly improves subsequent algorithm efficiency which requires $O(m^2)$ space and $O(q * m \log m)$ time complexity, where q is the number of query items considered. Eliminating rare items also avoids bizarre results when there is insufficient data for reliable statistical inference.[13]

Based on Amazon's recommender method,[15] co-occurrence statistics are precomputed from a random training set of 15 629 patients. This builds an item association matrix based on the definitions in Table 1, from which association statistics and Bayesian conditional probabilities in Table 2 can be estimated. If (repeat) items are counted, the results yield association metrics of support, confidence, and interest/lift.[28] Counting items only once per patient affords a natural interpretation of contingency statistics (e.g., relative ratio [RR], positive predictive value [PPV], baseline prevalence, Fisher's P-value) as in Table 3.

Generating and Evaluating Recommendations

To generate patient-specific recommendations, the recommender engine is queried with a patient's initial clinical items $(A_1, \ldots, A_q)$ to retrieve statistics from the precomputed association matrix for all possible target items $(B_1, \ldots, B_m)$ (excluding those already in the query set). Target items are ranked by a score such as ConditionalFreq$(B_j|A_i)(t)$, the maximum likelihood estimator for target item B_j occurring after query item A_i within time t.

Given q query items, q separate lists of scored target items are generated. We evaluated several approaches to aggregating these into a single result list. In example equation (1) below, we use N_{ABt} as our approximation for the average number of occurrences where target item B follows any of the $A_1, \ldots, A_q$ query items within time t. The optional weights $(1/N_{A_i})$ are inversely proportional to the query item baseline frequencies (lending more weight to less common, more specific query items):

$$\hat{N}_{ABt} = \frac{1}{\sum_{i=1}^{q} N_{A_i}} \sum_{i=1}^{q} \frac{1}{N_{A_i}} N_{A_i Bt} \qquad (1)$$

A naïve Bayes[29] aggregation approach makes the simplifying assumption that all query items occur independently of one another. This allows us to approximate the post-test probability in equation (2) with equation (3), which is estimated with the (co)-occurrence count statistics. Note that for recommendation *ranking* purposes, the denominator in (3) can be removed since it is constant for all target items B considered. Furthermore, it is often necessary to use the logarithm of the product series (i.e., sum of component logarithms) to avoid numerical underflow with extremely small values:

$$P(B \mid A_1, \ldots, A_q) = \frac{P(A_1, \ldots, A_q, B)}{P(A_1, \ldots, A_q)} = \frac{P(A_1, \ldots, A_q \mid B)}{P(A_1, \ldots, A_q)} P(B) \qquad (2)$$

$$\hat{P}(B \mid A_1, \ldots, A_q) = \frac{\prod_{i=1}^{q} P(A_i \mid B)}{\prod_{i=1}^{q} P(A_i)} P(B) \sim \frac{\prod_{i=1}^{q} \left(\frac{N_{A_i B}}{N_B}\right)}{\prod_{i=1}^{q} \left(\frac{N_{A_i}}{N}\right)} \left(\frac{N_B}{N}\right) \qquad (3)$$

Equations (4)–(8) outline another general approach to estimating the post-test probability of an item B given the item's pretest probability and the likelihood ratio for a single diagnostic test (i.e., occurrence of

Downloaded from http://jamia.oxfordjournals.org/ by guest on April 5, 2016

Figure 2: Recommender testing flow diagram. Each patient's data is modeled as a timeline of discrete clinical item instances as in the left table. For a set of validation patients, items from the first 4 h of each patient's timeline form a Query Set roughly corresponding to the patient's emergency room encounter. Each Query Set is fed into a Recommender module to produce a ranked Recommended List of all candidate orders not already included in the Query Set. The Recommended List is compared against the Validation Set of orders that actually occur within 24 h.

Item Date	Category	Description
12/25/2010 12:07	ECG	ECG 12-Lead
12/25/2010 12:13	Imaging	XR Chest 1V
12/25/2010 12:13	Lab	Blood Gases, Arterial
12/25/2010 12:13	Med (Intravenous)	Sodium Chloride (Intravenous)
12/25/2010 12:13	Microbiology	Blood Cultures
12/25/2010 12:13	PoCare Test	PoC Troponin I
12/25/2010 12:28	PoCare Test	PoC Venous Blood Gases and Lactate
12/25/2010 12:30	Lab Result	Albumin (Low)
12/25/2010 12:30	Lab Result	WBC (High)
12/25/2010 12:30	Lab Result	Sodium (Low)
12/25/2010 12:30	Lab Result	Blood Urea Nitrogen (High)
12/25/2010 12:30	Lab Result	Neutrophils, Absolute (High)
12/25/2010 12:30	Lab Result	WBC, Urine (Abnormal)
12/25/2010 12:30	Lab Result	Bacteria, Urine (Abnormal)
12/25/2010 12:30	Lab Result	Procalcitonin (High)
12/25/2010 12:32	Med (Intravenous)	Methylprednisolone (Intravenous)
12/25/2010 12:35	Med (Inhalation)	Levalbuterol (Inhalation)
12/25/2010 12:54	Med (Intravenous)	Levofloxacin (Intravenous)
12/25/2010 12:55	Med (Intravenous)	Piperacillin/Tazobactam (Intravenous)
12/25/2010 12:55	Med (Intravenous)	Vancomycin (Intravenous)
12/25/2010 13:36	PROBLEM_LIST	Pneumonia, organism unspecified
12/25/2010 13:36	ADMIT_DX	Shortness of breath
12/25/2010 13:36	Admission	Admit/Place Patient
12/25/2010 15:03	ECHO	ECHO – Transthoracic Echo +Doppler
12/25/2010 15:04	Microbiology	Urine Culture
12/25/2010 20:37	Imaging	US Vein Lower Extremity Bilateral r/o DVT
12/25/2010 20:37	Isolation	Droplet Isolation
12/25/2010 20:37	Lab	Lactic Acid
12/25/2010 20:37	Lab	Respiratory Virus DFA Panel-Reflex PCR
12/25/2010 20:37	Med (Inhalation)	Albuterol/Ipratropium (Inhalation)
12/25/2010 20:37	Med (Intravenous)	Pantoprazole (Intravenous)
12/25/2010 20:37	Med (Oral)	Acyclovir (Oral)
12/25/2010 20:37	Med (Oral)	Gabapentin (Oral)
12/25/2010 20:37	Med (Subcutaneous)	Heparin (Subcutaneous)
12/25/2010 20:37	Microbiology	Fungal Culture
12/25/2010 20:37	Microbiology	Respiratory Culture + Gram Stain
12/25/2010 20:37	Respiratory Care	RESP – Oxygen Administration
12/25/2010 22:29	Med (Intravenous)	Dexmedetomidine IV Infusion
12/25/2010 22:30	Respiratory Care	RESP – Non-Invasive Pos Pressure Vent (BIPAP)
12/25/2010 23:58	Lab	Heparin Activity Level
12/25/2010 23:58	Med (Intravenous)	Heparin (Intravenous)

Query Set

Validation Set

Recommender

Rank	Category	Description
1	Lab	NT - proBNP
2	Med (Inhalation)	Albuterol/Ipratropium (Inhalation)
3	Med (Intravenous)	Furosemide (Intravenous)
4	Med (Oral)	Aspirin (Oral)
5	Diet	Diet NPO
6	Lab	CK, MB (MASS)
7	ECHO	ECHO – Transthoracic Echo + Doppler
8	Med (Subcutaneous)	Heparin (Subcutaneous)
9	Med (Subcutaneous)	Enoxaparin (Subcutaneous)
10	Med (Oral)	Prednisone (Oral)

Recommended List

Table 1: Precomputed clinical item (co)-occurrence count statistics

Notation	Definition
n_A	Number of instances where item A occurs
n_{ABt}	Number of occurrences where item B follows item A within time threshold t
N_A	Number of patients for whom item A occurs
N_{ABt}	Number of patients for whom item B follows A within time t
N	Total number of patients

Lowercase values allow for repeat item counting, while uppercase values represent patient counts that ignore repeat items per patient.

an item A). Equations (4) – (6) include estimates based on (co)-occurrence count statistics:

$$\text{PreProb} = \text{Pretest probability} = P(B) \sim \frac{N_B}{N} \tag{4}$$

$$\text{PreOdds} = \text{Pretest odds} = \frac{\text{PreProb}}{1 - \text{PreProb}} \sim \frac{N_B}{N - N_B} \tag{5}$$

$$\text{LR}^+ = \text{Positive Likelihood Ratio} = \frac{\text{Sensitivity}}{1 - \text{Specificity}} = \frac{P(A|B)}{P(A|!B)} \sim \frac{N_{AB}/N_B}{N_A - N_{AB}/N - N_B} \tag{6}$$

$$\text{PostOdds} = \text{Post test odds} = \text{PreOdds} \times \text{LR}^+ \tag{7}$$

(Assumes positive test result. In this context, meaning that query item A occurred.)

$$\text{PostProb} = \text{Post test probability} = \frac{\text{PostOdds}}{1 + \text{PostOdds}} \tag{8}$$

A serial Bayes[30,31] aggregation approach uses the above equations with a similar simplifying assumption of independent query items. If the presence of each query item is considered a positive independent diagnostic test for the target item, the post-test probability from one test can be used as the pretest probability for the next test. This allows for the overall post-test odds for target item B to be calculated from a serial product of likelihood ratios as in equation (9):

$$\text{PostOdds} \sim \text{PreOdds} \times \prod_{i=1}^{q} \text{LR}_i^+ \sim \frac{N_B}{N - N_B} \times \prod_{i=1}^{q} \left(\frac{N_{A_i B}/N_B}{N_{A_i} - N_{A_i B}/N - N_B} \right) \tag{9}$$

While there is no well accepted notion of recommendation quality, accurately predicting subsequent items (Figure 2) is the most commonly

Downloaded from http://jamia.oxfordjournals.org/ by guest on April 5, 2016

Table 2: Association statistics and Bayesian probability estimates based on count statistics.

Metric	Related probability	Estimate	Notes
BaselineFreq	P(B)	n_B/N	Counting repeat items allows values >1.0. Interpreted as average number of times B occurs per patient.
"Support"	P(AB)	n_{ABt}/N	Not quite joint probability because n_{ABt} is a *directed* count where item A occurs *before* B, within time *t*.
ConditionalFreq "Confidence"	P(B|A) = P(AB) / P(A)	n_{ABt}/n_A	Counting repeat items allows values >1.0. Interpret as average number of times B occurs after A, within time *t*.
FreqRatio "Lift" "Interest"	P(B|A) * 1/P(B) = P(AB) / P(A)*P(B)	$(n_{ABt}/n_A)/(n_B/N)$	Equivalent to "TF*IDF."[29] Estimates likelihood ratio. Expect = 1 if A and B independent (i.e., P(B|A) = P(B)).
Prevalence Pre-test probability	P(B)	N_B/N	Only counts items once per patient. Interpret as percentage of all patients where item B occurs.
Positive predictive value (PPV) Post-test probability	P(B|A)	N_{ABt}/N_A	Percentage of patients where item B occurs after item A, within time *t*
Relative risk (RR)	P(B|A) / P(B|!A)	$(N_{ABt}/N_A)/((N_B-N_{ABt})/(N-N_A))$	Expect = 1 if A and B independent (i.e., P(B|A) = P(B) = P(B|!A)).

Table 3: Contingency 2×2 matrix of patient subgroup counts based on the occurrence of an item A or item B, from which association metrics like prevalence, positive predictive value (PPV), and relative risk (RR) can be estimated.

Subgroup	B	!B	Total
A	N_{ABt}	$N_A - N_{ABt}$	N_A
!A	$N_B - N_{ABt}$	$N - N_A - N_B + N_{ABt}$	$N - N_A$
Total	N_B	$N - N_B$	N

used metric that correlates with end-user satisfaction.[32] For a separate random test set of 1903 patients, we queried different recommender methods with each patient's first four hours of clinical items (average of 29) to produce a score-ranked list of all other candidate orders. The ranked recommended list is compared against the patient's actual subsequent orders within the first 24 h (average of 15) in terms of precision (positive predictive value) and recall (sensitivity) at 10 recommendations and receiver operating characteristic (ROC) analysis.

As previously noted,[13] standard accuracy measures reward recommendation of likely orders, though they may not make for "interesting" suggestions. To quantitatively recognize recommendations that are more specifically relevant to a query, we introduce alternative metrics of inverse frequency weighted precision and recall, based on the following indicator functions.

$TP_i = \{1$ *if recommended item i is a true positive, 0 if not*$\}$
$FP_i = \{1$ *if recommended item i is a false positive, 0 if not*$\}$
$FN_i = \{1$ *if recommended item i is a false negative, 0 if not*$\}$

Standard and inverse frequency weighted metrics are defined below for *k* recommendations, with the latter components weighted by the inverse baseline frequency of each item $i = (N_i/N)$. The common constant factor *N* will cancel out in the ratio to yield a weighting factor of $(1/N_i)$:

$$\text{Precision} = \frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k} TP_i + \sum_{i=1}^{k} FP_i}$$

$$\text{Weighted Precision} = \frac{\sum_{i=1}^{k} (1/N_i) TP_i}{\sum_{i=1}^{k} (1/N_i) TP_i + \sum_{i=1}^{k} (1/N_i) FP_i}$$

$$\text{Recall} = \frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k} TP_i + \sum_{i=1}^{k} FN_i}$$

$$\text{Weighted Recall} = \frac{\sum_{i=1}^{k} (1/N_i) TP_i}{\sum_{i=1}^{k} (1/N_i) TP_i + \sum_{i=1}^{k} (1/N_i) FN_i}$$

By "recommending" the probability of nonorder items such as patient death, need for ICU life support, hospital discharge, and readmission, the system can predict clinical outcomes. Doing so only requires outcome events to be labeled as another clinical item in the patient timeline. For intensive care unit (ICU) life support, a composite clinical item was defined as the occurrence of mechanical ventilation, vasopressor infusion, or continuous renal replacement therapy. Readmission was identified as a sequence of a single discharge order followed by a single admission order. For another random 1897 validation patients, the recommender was queried with the first 24 h of clinical items for the positive predictive value (post-test probability) of each outcome event within *t* time. Predicted probabilities were compared against actual events by ROC analysis. Validation patients were excluded if there were insufficient query items or the outcome already occurred within the query 24 h.

RESULTS

Tables 4–7 illustrate several example order recommendations. Figure 3 reports which recommender parameters optimize accuracy metrics. Table 8 reports the ROC area-under-curve (AUC) prediction accuracy for several clinical outcomes.

Downloaded from http://jamia.oxfordjournals.org/ by guest on April 5, 2016

Downloaded from http://jamia.oxfordjournals.org/ by guest on April 5, 2016

DISCUSSION

Analogous to commercial recommender systems, the described system recommends clinical orders and predicts clinical outcomes based on statistics data-mined from EMRs. Tables 4–7 qualitatively illustrate how ranking orders by maximum likelihood scores of PPV and ConditionalFreq ("Confidence") can identify likely subsequent orders, though these may not be "interesting" when they substantially overlap with the generic "best seller" list ranked by prevalence or BaselineFreq. Association measures like RR or FreqRatio ("Lift" or "Interest") rank less common but more specifically relevant items. Table 7 illustrates how patient-specific recommendations are refined with additional query information.

Results in Figure 3A demonstrate that personalizing order recommendations with the ConditionalFreq ranking improves prediction of real clinical order patterns compared to the BaselineFreq benchmark (27–35% precision at 10 recommendations, $P < 10^{-16}$) and similarly for PPV compared to prevalence (33–38%, $P < 10^{-10}$). With this level of predictive accuracy, it is not appropriate for such algorithms to "auto-pilot" medical decision making. The utility of such an approach must still be to provide relevant information to a *human* decision maker. In the context of passively providing suggestions to a human user however, these results appear useful when compared to the <33% average total click-through rate for top ten recommended links generated by Google AdWords[46] and <25% for related videos suggested on YouTube (interpolated by the empirically estimated Zipf function $Y = \beta X^{-\alpha}$ with $\beta = 5.6\%$ and $\alpha = 0.78$).[47]

The temporal relationship between clinical item instances is important to improve accuracy, by training the system with an item co-occurrence time threshold t matching the evaluation time frame. For example, when predicting orders occurring within 24 h of hospitalization, shorter time thresholds (e.g., 2 h) result in the system missing relevant associations outside the threshold, while longer time thresholds (e.g., any time) result in the system being distracted by associations outside the relevant evaluation period.

Table 4: Top orders by overall prevalence (pretest probability), the generic "best-seller" list.

Rank	Description	Prevalence (%)	Baseline Freq
1	Sodium chloride (IV)	90	3.97
2	CBC with differential	79	2.60
3	Metabolic panel, basic	77	2.65
4	Potassium chloride (IV)	68	1.76
5	Glucose (IV)	60	1.25
6	Docusate (Oral)	58	1.24
7	Prothrombin time (PT/INR (International Normalized Ratio))	58	1.93
8	Physical therapy evaluation	55	1.13
9	Metabolic panel, comprehensive	55	1.47
10	Diet NPO (Non per os)	53	1.11

BaselineFreq counts (repeat) items instead of patients. For example, an average of 2.6 orders for CBC (Complete Blood Count) with Differential occurs for all patients, with 79% of patients receiving them.

Methods based on patient counts (i.e., PPV and prevalence) slightly outperform those based on item counts (i.e., ConditionalFreq, BaselineFreq), as they may be less confounded by orders that are used repeatedly for a smaller number of patients. Results from Figure 3A and Table 9 show a slight but consistent improvement in prediction accuracy when weighting query items to favor the influence of less common, more specific items. Despite a more theoretically sound basis, naïve Bayes and serial Bayes aggregation methods require greater implementation complexity yet yield no significant accuracy benefit over weighted averaging.

Results from related research generally cannot be compared directly, given different use cases and problem space. Early work illustrated the concept of identifying clinical item associations, but concluded with only descriptive findings of the top associations found.[16,19] Subsequent efforts to predict clinical orders with such methods focused on problem spaces with dozens of possible candidate items.[21–23] The problem space in this manuscript consists of *hundreds* of candidate orders, resulting in substantially different problem "difficulty" and expected accuracy.[48] Recommendations based on baseline prevalence provide a more consistent internal benchmark, which is also notably much more accurate than random. Prior research to impute missing medications confirms that baseline prevalence can be a difficult benchmark to surpass, with more complex nearest neighbor and regularized logistic regression models not necessarily yielding meaningful improvement.[22] Association rule methods explored here do modestly improve accuracy, though there may be substantial opportunity for further improvement. Research into Bayesian networks appears to improve clinical order prediction accuracy over association rules,[21] but the tradeoff in computational complexity appears to require a smaller problem space. Perhaps more importantly, traditional metrics of accuracy may not even reflect the most useful recommendations.

We introduced alternative metrics, the inverted frequency weighted precision and recall, to credit the more compelling prediction of uncommon but "interesting" items (e.g., rifaximin) over common but mundane items (e.g., CBC). Interestingly, Figure 3C indicates that PPV and ConditionalFreq methods still achieve the best weighted precision. Weighted recall is where RR and FreqRatio based methods show substantial improvement (4–16%, $P < 10^{-16}$) compared to baseline prevalence or PPV-based predictions. This reinforces the notion that the two approaches satisfy different goals (predicting likely events vs finding "interesting" suggestions).

Table 9 reports the system's ability to predict clinical outcomes, with only a semantic difference between "predicting outcomes" and "recommending orders." The system effectively acts as a naïve Bayes classifier for individual outcomes, using prior clinical items as features to generate a PPV score. The system yields a prediction ROC AUC (c-statistic) of 0.84 for 30 day mortality, 0.84 for 1 week need for ICU life support, 0.80 for 1 week discharge from the hospital, and 0.68 for hospital readmission. Though different patient populations and evaluation periods prevent direct comparison, these accuracies are on par with state-of-the art prognosis scoring systems. Inverted queries that "recommend" items *preceding* a clinical outcome event[14] can also effectively act as a tool for feature selection to infer risk factors.

Several limitations in this work will require future study. The primary concern with learning practice patterns from historical data is that it will favor *common* practices that are not necessarily "correct." Such a system could theoretically recommend inappropriate orders, resulting in a positive feedback loop that continually reinforces poor decision making into common behavior. While *some* clinicians will certainly enter inappropriate orders *sometimes*, Condorcet's Jury

Table 5: Top orders occurring within 24 h of an order for Spironolactone (Oral), a diuretic commonly used in the management of heart failure and liver cirrhosis.

Rank	Description	PPV (%)	Prevalence (%)	Conditional Freq	Baseline Freq
1	Metabolic panel, basic	42	77	0.41	2.65
2	Furosemide (Oral)	41	11	0.52	0.31
3	Magnesium, serum	35	52	0.35	1.84
4	Prothrombin time (PT/INR)	35	58	0.35	1.93
5	CBC with differential	31	79	0.28	2.60
6	Potassium chloride (Oral)	28	27	0.34	0.82
7	Furosemide (IV)	27	17	0.35	0.84
8	Pantoprazole (Oral)	24	38	0.19	0.62
9	CBC	24	53	0.20	1.50
10	Sodium chloride (IV)	23	90	0.19	3.97

Results are ranked by PPV (positive predictive value $\sim$ post-test probability), meaning 41% of patients subsequently receive an order for Furosemide (Oral) (another common diuretic), with an average of 0.52 orders per patient (ConditionalFreq = "Confidence"). Related orders include additional diuretics (Furosemide), monitoring of electrolytes affected by diuresis (Magnesium, Serum), and repletion of lost electrolytes (Potassium Chloride). Notably, many of the orders are the same "best-seller" items that are common overall.

Table 6: Top orders occurring within 24 h of Spironolactone (Oral), ranked by Fisher's exact test *P*-value, where RR (relative risk)>1.

Rank	Description	P-Fisher	RR	PPV (%)	Prevalence (%)	Freq Ratio	Conditional Freq	Baseline Freq
1	Furosemide (Oral)	6.E-87	4.0	41	11	1.7	0.52	0.31
2	Carvedilol (Oral)	2.E-27	3.1	19	7	1.0	0.19	0.18
3	Digoxin (Oral)	1.E-25	4.4	12	3	1.8	0.12	0.07
4	Rifaximin (Oral)	1.E-16	5.5	6	1	1.9	0.07	0.03
5	Furosemide (IV)	1.E-11	1.7	27	17	0.4	0.35	0.84
6	Sildenafil (Oral)	1.E-10	5.4	4	1	1.6	0.04	0.03
7	Propranolol (Oral)	7.E-10	3.6	5	2	1.4	0.05	0.04
8	Lactulose (Oral)	7.E-09	2.4	8	4	1.2	0.12	0.10
9	PoC (Point of Care) Venous Blood Panel	3.E-08	3.9	4	1	0.8	0.11	0.14
10	Diet Sodium Restricted	3.E-08	1.6	20	13	0.6	0.15	0.24

Furosemide (Oral) is 4.0 times more likely to be ordered after Spironolactone than if Spironolactone was not ordered. This correlates to the FreqRatio = "Lift" = "Interest" indicating Furosemide (Oral) is ordered 1.7 times more often after Spironolactone than for all patients. Even without a clinical diagnosis or patient history, this example illustrates how a clinical order (i.e., spironolactone) is itself predictive of other orders specifically relevant to the implied clinical scenarios. For example, furosemide, carvedilol, digoxin, and sodium restriction are used to manage congestive heart failure,[33] while furosemide, rifaximin, propranolol, and lactulose all help manage complications of liver cirrhosis (ascites,[34] hepatic encephalopathy,[35] and esophageal varices[36]). A less obvious suggestion is sildenafil, likely based on its concurrent use for pulmonary hypertension.[37]

Theorem[49] posits that aggregating the nonrandom decisions of many converges towards correctness. This is the same basis argument behind the "wisdom-of-the-crowd[50]" and for machine-learning boosting algorithms that generate strong classifiers from individually weak ones.[51] The ability to predict clinical outcomes offers a tempting possibility to link recommendations to favorable outcomes, though this would almost certainly be undermined by patient confounders without effective cohort balancing.[52] We are exploring additional work comparing recommendations against clinical practice guidelines,[53] but ultimately this concern will only be proven or disproven in a prospective clinical trial. Towards that end, we are developing human-computer interface prototypes of the system for simulation testing with real clinicians. This will be necessary to provide more specific evaluation of click through rates, user satisfaction with generated recommendations, and impacts on clinical decision making. In the meantime, the methods developed here can be useful for applications

Downloaded from http://jamia.oxfordjournals.org/ by guest on April 5, 2016

Table 7: Top orders occurring within 24 h of an order for Spironolactone (Oral) *and* Lactulose (Oral), score-ranked by *P*-value, where RR > 1 and scores are estimated by weighted averaging of underlying count statistics.

Rank	Description	P-Fisher	RR	PPV (%)	Prevalence (%)	Freq Ratio	Conditional Freq	Baseline Freq
1	Rifaximin (Oral)	2.E-54	14.2	13	1	5.1	0.17	0.03
2	Furosemide (Oral)	2.E-21	2.3	25	11	1.1	0.33	0.31
3	Propranolol (Oral)	5.E-12	4.2	6	2	1.8	0.07	0.04
4	Zinc Sulfate (Oral)	2.E-07	2.7	6	2	2.1	0.09	0.04
5	Albumin, Fluid	4.E-07	3.6	4	1	1.9	0.06	0.03
6	Digoxin (Oral)	7.E-05	2.1	6	3	0.9	0.06	0.07
7	Alpha Fetoprotein	1.E-04	3.2	3	1	1.8	0.02	0.01
8	Ammonia	3.E-04	1.9	6	4	0.9	0.07	0.08
9	Lactulose (Enema)	5.E-04	2.9	3	1	2.5	0.05	0.02
10	Sildenafil (Oral)	5.E-04	3.1	2	1	1.0	0.03	0.03

The results illustrate progressively patient focused suggestions by demoting orders for non-liver diseases while promoting rifaximin, zinc sulfate, lactulose enemas, and ammonia levels for hepatic encephalopathy; furosemide, propranolol, and albumin fluid checks for portal hypertension with ascites; and alpha fetoprotein for monitoring hepatocellular carcinoma.

Table 8: ROC AUC (c-statistic) prediction results for clinical outcomes based on 1897 validation patients' first 24 h of query clinical items.

Property	Death	ICU Life Support			Hospital Discharge / Length of Stay				Readmission
Evaluation period	30 days	2 days	4 days	1 week	2 days	4 days	1 week	2 week	30 days
Patients included	1890	1772	1772	1772	1742	1742	1742	1742	1890
Patients w/outcome, *n* (%)	35 (1.9)	17 (1.0)	28 (1.6)	41 (2.3)	387 (22.2)	1020 (58.6)	1380 (79.2)	1570 (90.1)	157 (8.3)
ROC AUC (c-stat)	0.84	0.87	0.83	0.84	0.69	0.75	0.80	0.82	0.68
ROC AUC 95% CI	0.75-0.94	0.80-0.95	0.76-0.90	0.79- 0.89	0.66-0.72	0.72-0.77	0.78-0.83	0.79-0.85	0.64-0.73
Related Prediction Models	APACHE, MPM, SAPS[38]	CURB-65, PSI, SCAP, REA-ICU[39]			Tu et al.[40], ISS, NISS,[41] CSI[42]				Amarasingham et al.,[43], LACE,[44] CMS[45]
ROC AUC for Related	0.75-0.90	0.69-0.81			0.69 (Only Tu et al. reports as ROC AUC)				0.56-0.72

Confidence intervals (95%) empirically estimated by bootstrapping 1000 data samples with replacement. ROC = Receiver Operating Characteristic; AUC = Area Under Curve; CI = Confidence Interval; APACHE = Acute Physiology and Chronic Health Evaluation; MPM = Mortality Probability Models; SAPS = Simplified Acute Physiology Score; CURB-65 = Confusion, Urea, Respiratory Rate, Blood Pressure, 65 Years Old; PSI = Pneumonia Severity Index; SCAP = Severity Community-Acquired Pneumonia; REA-ICU = Risk of Early Admission to ICU; ISS = Injury Severity Score; NISS = New Injury Severity Score; CSI = Computerized Severity Index; LACE = Length of Stay, Acuity, Comorbidity, Emergency Department Use; CMS = Center for Medicare and Medicaid Services

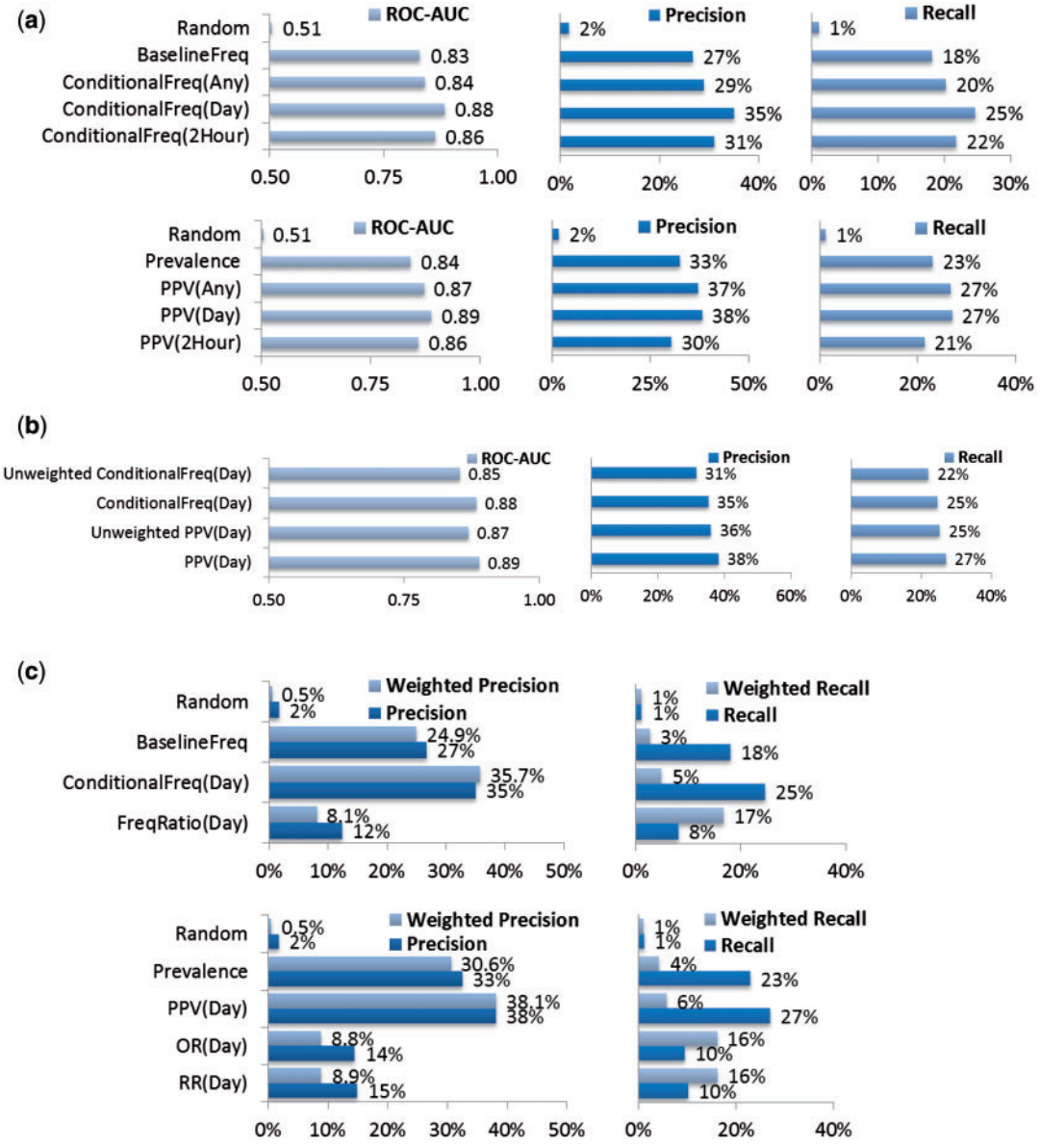
ranging from patient risk stratification to practice pattern analysis to computer-aided drafting of decision support rules.

CONCLUSIONS

Collaborative filtering methods successful in non-biomedical domains can automatically generate CDS from historical practice data in the form of order recommendations. Clinical orders can be both results and predictive features for other clinical decisions and outcomes. Association-based recommendations are predictive of real practice patterns and clinical outcomes as compared to baseline benchmarks. Temporal relationships are important to improve accuracy. Different evaluation metrics will satisfy different query goals to either predict likely events or find "interesting" suggestions. This work represents another step towards real-time CDS tools that will unlock the Big Data potential of EMRs, uniquely tying together the afferent and efferent limbs of a learning health system.

RESEARCH AND APPLICATIONS Downloaded from http://jamia.oxfordjournals.org/ by guest on April 5, 2016

Figure 3: (A) – Accuracy of clinical order recommendations using different parameters. For 1903 validation patients, clinical items from the first 4 h of their hospital encounter are used to query for a ranked list of clinical order recommendations. Each recommended list is validated against the actual set of new orders occurring within 24 h of hospitalization. Metrics include average receiver operating characteristic area under curve (ROC AUC) (c-statistic) and top 10 recommendation precision (positive predictive value) and recall (sensitivity). Ranking by BaselineFreq serves as a reference benchmark. ConditionalFreq methods are further refined by what time threshold t was used when counting item co-occurrences. The Prevalence (pretest probability) and PPV (positive-predictive value $\sim$ post-test probability) methods are directly analogous to BaselineFreq and ConditionalFreq, except they count *patients* with item co-occurrences, ignoring repeat items. Two-tailed, paired t-tests for the first row results compared to the BaselineFreq benchmark all yield $P < 10^{-16}$, while second row results compared to the Prevalence benchmark all yield $P < 10^{-10}$. (B) Accuracy of clinical order recommendations/predictions using simple unweighted averaging vs the default weighted averaging of underlying count statistics to favor the influence of less common, more specific query items. Two-tailed, paired t-tests of respective unweighted vs weighted aggregation methods all yield $P < 10^{-27}$. (C) Inverse frequency weighted accuracy of clinical order recommendations, all using a time threshold $t = 1$ day and considering the top 10 recommendations. Two-tailed, paired t-tests for first row results compared to the ConditionalFreq(Day) method all yield $P < 10^{-29}$, while second row results compared to the PPV(Day) method all yield $P < 10^{-16}$.



Downloaded from http://jamia.oxfordjournals.org/ by guest on April 5, 2016

Table 9: Prediction accuracy for 30-day mortality using different score aggregation methods.

Aggregation Method	ROC AUC	ROC AUC 95% CI
Unweighted ConditionalFreq	0.82	0.73-0.91
Unweighted PPV	0.84	0.74-0.93
Weighted ConditionalFreq	0.84	0.75-0.93
Weighted PPV	0.84	0.75-0.94
Naïve Bayes	0.82	0.74-0.90
Serial Bayes	0.82	0.74-0.90

Weighted average methods perform best, though the differences are incremental. Methods that may appear more statistically sound (Naïve Bayes, Serial Bayes) perform no better, limiting the value of their increased implementation complexity.

FUNDING

This work was supported by the Stanford Translational Research and Applied Medicine program in the Department of Medicine and the Stanford Learning Healthcare Systems Innovation Fund and the Stanford Clinical and Translational Science Award (CTSA) to Spectrum (UL1 TR001085). The CTSA program is led by the National Center for Advancing Translational Sciences at the National Institutes of Health (NIH).

J.H.C supported in part by VA Office of Academic Affiliations and Health Services Research and Development Service Research funds.

R.B.A. is supported by NIH/National Institute of General Medical Sciences PharmGKB resource, R24GM61374, as well as LM05652 and GM102365. Additional support is from the Stanford NIH/National Center for Research Resources CTSA award number UL1 RR025744.

Patient data extracted and de-identified by Stanford Translational Research Integrated Database Environment (STRIDE) project, a research and development project at Stanford University to create a standards-based informatics platform supporting clinical and translational research. The STRIDE project described was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through grant UL1 RR025744.

Content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, VA, or Stanford Healthcare.

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

J.H.C. conceived of the study, implemented the algorithms, performed the analysis, and drafted the initial manuscript. T.P. extracted and normalized the contents of the electronic health record data. R.B.A. contributed to algorithm design, analysis design, manuscript revisions, and supervised the study.

REFERENCES

1. Richardson WC, Berwick DM, Bisgard JC, *et al. Crossing the Quality Chasm: A New Health System for the 21st Century.* Institute of Medicine, Committee on Quality of Health Care in America Committee on Quality of Health Care in America; 2001. Washington DC: National Academy Press.
2. Lauer MS, Bonds D. Eliminating the 'expensive' adjective for clinical trials. *Am Heart J,* 2014;167(4):419–420.
3. Office of the National Coordinator for Health Information Technology (ONC). Health information technology: standards, implementation specifications, and certification criteria for electronic health record technology, 2014 edition; revisions to the permanent certification program for health information technology. Final rule. *Fed. Regist.* 2012;77(171):54163–54292.
4. Longhurst CA, Harrington RA, Shah NH. A 'green button' for using aggregate patient data at the point of care. *Health Aff.* 2014;33(7):1229–1235.
5. Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. *N Engl J Med.* 2011;365(19):1758–1759.
6. Smith M, Saunders R, Stuckhardt L, McGinnis JM. *Best Care at Lower Cost: the Path to Continuously Learning Health Care in America.* Institute of Medicine, Committee on the Learning Health Care System in America; 2012. Vol. 6, pp. 149–188. Washington DC: National Academy Press.
7. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Aff.* 2014;33(7):1163–1170.
8. de Lissovoy G. Big data meets the electronic medical record: a commentary on 'identifying patients at increased risk for unplanned readmission'. *Med Care.* 2013;51(9):759–760.
9. Kaushal R, Shojania KG, Bates DW. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. *Arch Intern Med.* 2003;163(12):1409–1416.
10. Overhage J, Tierney W. A randomized trial of 'corollary orders' to prevent errors of omission. *JAMIA.* 1997;4(5):364–375.
11. Bates DW, Kuperman GJ, Wang S, *et al.* Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *JAMIA.* 2003;10(6):523–530.
12. Sittig DF, Wright A, Osheroff JA, *et al.* Grand challenges in clinical decision support. *J Biomed Inform.* 2008;41(2):387–392.
13. Chen JH, Altman RB. Mining for clinical expertise in (undocumented) order sets to power an order suggestion system. *AMIA Summits Transl Sci Proc.* 2013;2013:34–38.
14. Chen JH, Altman RB. Automated physician order recommendations and outcome predictions by data-mining electronic medical records. *AMIA Summits Transl Sci Proc.* 2014:206–210.
15. Linden G, Smith B, York J. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Comput.* 2003;7(1):76–80.
16. Doddi S, Marathe A, Ravi SS, Torney DC. Discovery of association rules in medical data. *Med Inform Internet Med.* 2001;26(1):25–33.
17. Klann J, Schadow G, Downs SM. A method to compute treatment suggestions from local order entry data. *AMIA Annu Symp Proc.* 2010;2010:387–391.
18. Klann J, Schadow G, McCoy JM. A recommendation algorithm for automating corollary order generation. *AMIA Annu Symp Proc.* 2009;2009:333–337.
19. Wright A, Sittig DF. Automated development of order sets and corollary orders by data mining in an ambulatory computerized physician order entry system. *AMIA Annu Symp Proc.* 2006;2006:819–823.
20. Zhang Y, Padman R, Levin JE. Paving the COWpath: data-driven design of pediatric order sets. *JAMIA.* 2014;21(e2):e304–e311.
21. Klann JG, Szolovits P, Downs SM, Schadow G. Decision support from local data: creating adaptive order menus from past clinician behavior. J Biomed Inform. 2014;48:84–93.
22. Hasan S, Duncan GT, Neill DB, Padman R. Automatic detection of omissions in medication lists. *JAMIA.* 2011;18(4):449–458.
23. Wright AP, Wright AT, McCoy AB, and Sittig DF. The use of sequential pattern mining to predict next prescribed medications. *J Biomed Inform.* 2014;53:73–80.
24. Zhang Y, Levin JE, Padman R. Data-driven order set generation and evaluation in the pediatric environment. *AMIA Annu Symp Proc.* 2012;2012:1469–1478.
25. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE–An integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc.* 2009;2009:391–395.
26. Hernandez P, Podchiyska T, Weber S, Ferris T, Lowe H. Automated mapping of pharmacy orders from two electronic health record systems to RxNorm within the STRIDE clinical data warehouse. *AMIA Annu Symp Proc.* 2009;2009(2):244–248.
27. Wright A, Bates DW. Distribution of problems, medications and lab results in electronic health records: the pareto principle at work. *Appl Clin Inform.* 2010;1(1):32–37.
28. Brin S, Motwani R, Ullman JD, Tsur S. Dynamic itemset counting and implication rules for market basket data. In *SIGMOD 1997, Proceedings ACM*

RESEARCH AND APPLICATIONS

Downloaded from http://jamia.oxfordjournals.org/ by guest on April 5, 2016

SIGMOD International Conference on Management of Data. May 13-15, 1997, Tucson, Arizona, USA, 1997: 255–264.
29. Manning CD, Schütze H. *Foundations of Statistical Natural Language Processing*, Cambridge, Massachusetts: MIT Press;1999.
30. Morgan AA, Chen R, Butte AJ. Likelihood ratios for genome medicine. *Genome Med.* 2010;2(5):30.
31. Musen MA, Middleton B, Greenes RA. In: Shortliffe E, Cimino J, eds. Clinical Decision-Support Systems in Biomedical Informatics. London: Springer-Verlag; 2014(22):643–674.
32. Shani G, Gunawardana A. Evaluating recommendation systems. *Recomm Syst Handb.* 2011;12(19):1–41.
33. Yancy CW, Jessup M, Bozkurt B, *et al*. ACCF/AHA guideline for the management of heart failure: A report of the american college of cardiology foundation/american heart association task force on practice guidelines," *Circulation. 2013*;128:240–327.
34. Runyon BA. AASLD PRACTICE GUIDELINE Management of Adult Patients with Ascites Due to Cirrhosis: Update 2012. Hepatology. 2013;57:1651–1653.
35. Vilstrup H, Wong P. Hepatic encephalopathy in chronic liver disease. Pract Guidel by AASLD EASL. 2014;2014:1–74.
36. LaBrecque D, Khan A, Sarin S, Le Mair A. Esophageal varices. *World Gastroenterol Organ Glob Guidel.* 2014;2014:1–14.
37. Gali N, Hoeper MM, Humbert M, *et al*. Guidelines for the diagnosis and treatment of pulmonary hypertension. *Eur Heart J.* 2009;30: 2493–2537.
38. Lemeshow S, Le Gall JR. Modeling the severity of illness of ICU patients. *A systems update*. JAMA. 1994;272(13):1049–1055.
39. Renaud B, Labarère J, Coma E, *et al*. Risk stratification of early admission to the intensive care unit of patients with no major criteria of severe community-acquired pneumonia: development of an international prediction rule. *Crit Care.* 2009;13(2):R54.
40. Tu JV, Jaglal SB, Naylor CD. Multicenter validation of a risk index for mortality, intensive care unit stay, and overall hospital length of stay after cardiac surgery. Steering Committee of the Provincial Adult Cardiac Care Network of Ontario. *Circulation.* 1995;91:677–684.
41. Lavoie A, Moore L, LeSage N, Liberman M, Sampalis JS. The injury severity score or the new injury severity score for predicting intensive care unit admission and hospital length of stay? *Injury* 2005;36:477–483.
42. Horn SD, Sharkey PD, Buckle JM, Backofen JE, Averill RF, Horn RA. The relationship between severity of illness and hospital length of stay and mortality. *Med Care.* 1991;29(4):305–317.
43. Amarasingham R, Moore BJ, Tabak YP, *et al*. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Med Care.* 2010;48(11):981–988.
44. van Walraven C, Dhalla IA, Bell C, *et al*. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *CMAJ.* 2010;182(6):551–557.
45. Kansagara D, Englander H, Salanitro A, *et al*. Risk prediction models for hospital readmission: a systematic review. *JAMA.* 2011;306(15):1688–1698.
46. Divecha F. *Google AdWords Click Through Rates Per Position*. 2009. [Online]. http://www.accuracast.com/articles/advertising/adwords-click through/. Accessed June 29, 2015.
47. Zhou R, Khemmarat S, Gao L. The impact of YouTube recommendation system on video views. *Proc 10th ACM SIGCOMM Conf Internet Meas*. 2010;404–410.
48. Schröder G, Thiele M, Lehner W. Setting goals and choosing metrics for recommender system evaluations. *In* CEUR Workshop Proc. 2011;811:78–85.
49. Austen-Smith D, Banks JS. Information aggregation, rationality, and the Condorcet jury theorem. *Am Polit Sci Rev.* 1996;90(1):34–45.
50. Surowiecki J. *The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisom Shapes Business, Economies, Societies, and Nations*. 2004, Doubleday, New York.
51. Schapire RRE. The strength of weak learnability. *Mach Learn.* 1990;227:197–227.
52. Leeper NJ, Bauer-Mehren A, Iyer SV, Lependu P, Olson C, Shah NH. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. *PLoS One.* 2013;8(5):e63499.
53. Chen JH, Altman RB. Data-mining electronic medical records for clinical order recommendations: wisdom of the crowd or tyranny of the mob? *AMIA Summits Transl Sci Proc.* 2015;2015:435–439.

AUTHOR AFFILIATIONS

[1]Center for Innovation to Implementation (Ci2i), Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, USA

[2]Center for Primary Care and Outcomes Research (PCOR), Stanford University, Stanford, CA, USA

[3]Biomedical Informatics Training Program, Stanford University, Stanford, CA, USA

[4]Departments of Bioengineering, Genetics, and Medicine, Stanford University, Stanford, CA, USA

Downloaded from http://jamia.oxfordjournals.org/ by guest on April 5, 2016