Automated Physician Order Recommendations and Outcome Predictions by Data-Mining Electronic Medical Records

Jonathan H. Chen, MD, PhD[1], Russ B. Altman, MD, PhD[1,2*]
1 Department of Medicine, Stanford University, Stanford, CA 94305, USA.
2 Departments of Bioengineering and Genetics, Stanford University, Stanford, CA 94305, USA.
*To whom correspondence should be addressed. E-mail: russ.altman@stanford.edu

Abstract

The meaningful use of electronic medical records (EMR) will come from effective clinical decision support (CDS) applied to physician orders, the concrete manifestation of clinical decision making. CDS development is currently limited by a top-down approach, requiring manual production and limited user awareness. A statistical data-mining alternative automatically extracts expertise as association statistics from structured EMR data (>5.4M data elements from >19K inpatient encounters). This powers an order recommendation system analogous to commercial systems (e.g., Amazon.com's "Customers who bought this…"). Compared to a standard benchmark, the association method improves order prediction precision from 26% to 37% ($p<0.01$). Introducing an inverse frequency weighted recall metric demonstrates a quantifiable improvement from 3% to 17% ($p<0.01$) in recommending more specifically relevant orders. The system also predicts clinical outcomes, such as 30 day mortality and 1 week ICU intervention, with ROC AUC of 0.88 and 0.78 respectively, comparable to state-of-the-art prognosis scores.

Introduction

Electronic medical records (EMR) can improve patient safety and healthcare cost efficiency, but that depends on meaningful use of the data[1]. This will require effective clinical decision support (CDS) content, particularly to drive clinical orders (labs, imaging, medications, etc.), the concrete manifestation of clinical decision making. Order sets, risk scores, and similar CDS constructs help reinforce consistency and compliance with best-practices[2,3], but their conventional development is limited by a top-down approach. This approach requires manual production of CDS content, feasible for only a limited number of common scenarios, and often with limited end-user awareness[4]. With the progressive digitization of clinical data in EMRs, a Big Data[5,6] approach can instead crowd-source clinical expertise from the bottom-up by data-mining EMRs. Such an approach could continuously "learn" in real-time by streaming in accumulating EMRecords into data-driven models of clinical expertise, even as it is simultaneously applied to patient care with direct EMR integration.

Background

Prior work in automated CDS content development includes association rules and Bayesian networks between orders and diagnoses, and review of possible order set and corollary order content by subject experts[7–10]. With inspiration from analogous problems of information retrieval in recommender systems, collaborative filtering, market basket analysis, and natural language processing, we initiated an item association order recommendation framework[11] analogous to Netflix or Amazon.com's "Customer's who bought A also bought B" system[12]. Here we update our initial efforts with a much larger dataset that includes non-order data to better define a patient's clinical context, propose an alternative evaluation metric to identify recommendation methods that highlight items specifically relevant to a given clinical scenario, and use the framework to predict clinical outcomes.

Methods

Deidentified, structured patient data from inpatient hospitalizations at Stanford University Hospital in 2011 was extracted by the STRIDE project[13]. Extracted data covers patient encounters starting from their initial (emergency room) presentation until hospital discharge. With >19K distinct patients, the data consists of >5.4M instances of >17K distinct clinical items, with patients, instances, and items respectively analogous to documents, words, and vocabulary items. The clinical items include >3,500 medication, >1,000 laboratory, >800 imaging, and >700 nursing orders. Non-order items include >1,000 lab results, >5,800 problem list entries, >3,400 admission diagnosis ICD9 codes, and patient demographics on age, gender, and date of death. Numerical data was binned into categorical data, particularly lab results, based on "abnormal" flags as established by the clinical laboratory. The ICD9 coding hierarchy was collapsed as necessary into diagnosis codes with a significant number of instances.

The relationship between item instances covered and the top clinical items considered is consistent with the "80/20 rule" in the form of a power law distribution[14]. This property allows one to ignore most clinical items with minimal information loss, in this case by ignoring sparsely populated clinical items with <256 instances (0.005% of

all instances), reducing the effective item count from >17K to 1.5K (9%), while only reducing coverage of item instances from 5.4M to 5.1M (94%). Computational efficiency of subsequent order recommendations improves significantly with this simplification, given methods requiring $O(m^2)$ space and $O(q * m \log m)$ time complexity, where m is the number of distinct clinical items considered and q is the number of query items for a specific recommendation.

A pre-computation step collects frequency statistics on clinical item instance co-occurrences from a training set of 16,408 randomly selected patients to build an item association matrix, based on the definitions in Table 1. These statistics are the basis for subsequent order recommendations by approximating Bayesian conditional probabilities as in Table 2.

Notation	Definition
$\mathbf{n_A}$	Number of occurrences of order A
$\mathbf{n_{ABt}}$	Number of occurrences of order B following an order A within time t
N	Total number of patients

Table 1 - Pre-computed frequency statistics for clinical items. Counting repeats allowed.

Probability	Estimate	Notation / Notes
P(A)	nA / N	baselineFreq(A)
P(AB)	nAB / N	nAB ("Support") only counts directed association where A occurs *before* B
P(B|A) = P(AB) / P(A)	nAB / nA	conditionalFreq(B|A) ("Confidence") Frequency of B, given A
P(B|A) / P(B) = P(AB) / P(A)*P(B)	(nAB/nA) / (nB/N)	freqRatio(B|A). Estimates likelihood ratio. Expect = 1, if A and B occur independently

Table 2 - Bayesian probability estimates based on item frequency statistics.

To generate order recommendations from the above association statistics, query clinical items ($A_1,\ldots,A_q$) are used to select item association pairs from the pre-computed association matrix for all possible target orders ($B_1,\ldots,B_m$). Target orders are ranked by a score such as conditionalFreq($B_j|A_i$), the maximum likelihood estimator for the probability of order B_j occurring after query item A_i. As previously noted[11], ranking by conditionalFreq identifies likely orders, but also tends to yield non-specific orders (e.g., CBC, IV saline) that are common overall, yet not necessarily "interesting." To identify orders more significantly relevant to the query, recommendations are ranked or filtered by freqRatio(B|A), comparable to the TF*IDF (term frequency * inverse document frequency) natural language processing concept[15].

To quantify the significance of item associations, *-2 log freqRatio* can approximate a chi-square statistic[15] or the chi-square statistic can be directly calculated by comparing observed vs. expected pre-computed occurrence counts. Issues with misinterpreting association strengths in the setting of inadequate data (heuristics advise at least 5 occurrences to be reliable[15]), are mitigated by excluding rare items occurring <0.005% of the time as previously described.

Given q query items, the above method generates q scored lists of all m possible orders. These are aggregated into a single scored recommendation list by taking a weighted average of the component scores, weighted inversely proportional to their respective query item baseline frequencies (lending more weight to less common, more specific query items). Unweighted score averaging and a Naïve Bayes[15] style composite product of the component conditional probabilities (i.e., conditional frequencies) were also attempted, though the weighted average method was retained as it yielded the best results.

While there is no well accepted notion of recommendation quality, accuracy in predicting subsequent items is the most commonly measured, with precision (positive predictive value) and recall (sensitivity) correlating with end-user satisfaction[16]. A test set of 1,903 patients was randomly selected, separate from the training set. For each test patient, all clinical items from the first 4 hours of their hospital encounter were used (average of 29) to query for 10 recommended orders that were compared against the actual subsequent orders within the first 24 hours (average of 15). To quantitatively recognize recommenders that yield results that are more meaningfully relevant to a query and not simply common, we introduce the alternative metrics of inverse frequency weighted precision and recall, based on the following function definition: *TP(i) = {1 if recommended item i is a true positive, 0 if not}*. Likewise *FP(i)* for false positives and *FN(i)* for false negatives. The inverse frequency weighted precision and recall metrics are defined below in summation notation, with components weighted by the inverse baseline frequency of each item *i* (n_i/N). Note that the common constant factor N can be cancelled out to yield:

$$\text{Weighted Precision} = \Sigma\,(1/n_i)*TP(i) \,/\, (\,\Sigma\,(1/n_i)*TP(i) + \Sigma\,(1/n_i)*FP(i)\,)$$

$$\text{Weighted Recall} = \Sigma\,(1/n_i)*TP(i) \,/\, (\,\Sigma\,(1/n_i)*TP(i) + \Sigma\,(1/n_i)*FN(i)\,)$$

The association framework was also applied towards "recommending" non-order items to predict outcomes such as patient death and ICU intervention. For the latter, a composite "AnyICU" clinical item was defined as the occurrence of interventions including mechanical ventilation, vasopressor infusion (epinephrine, norepinephrine, dopamine, phenylephrine, vasopressin, dobutamine), or continuous renal replacement therapy (CRRT). Taking 1,905 test patients separate from the training set, their first 24 hours of clinical items were used to query the association model to score the probability (conditionalFreq(B|A)$_t$) of an outcome event within *t* time (30 days for death, 1 week for AnyICU) and compared them vs. actual event rates by receiver operating characteristic (ROC) analysis.

Results

Table 3 illustrates example order recommendations. Table 4 reports accuracy metrics for different recommendation methods, illustrating the trends toward the best results. Table 5 illustrates an inverted query example, identifying items commonly *preceding* an outcome event. Table 6 reports the ROC area-under-curve (AUC) prediction accuracy for outcomes of 30 day mortality and 1 week use of AnyICU.

		Frequency / Likelihood			
Rank	**Description**	**Conditional**	**Baseline**	**Ratio**	**p**
1	TYPE AND SCREEN	0.98	0.78	1.3	0.00
2	Pantoprazole (Intravenous)	0.75	0.42	1.8	0.00
3	TRANSFUSE RBC	0.55	0.52	1.1	0.20
4	PANTOPRAZOLE IV INFUSION	0.51	0.03	16.0	0.00
5	CONSULT MEDICINE	0.32	0.16	2.0	0.00
6	LIPASE	0.29	0.26	1.1	0.15
7	ISTAT TROPONIN I	0.28	0.28	1.0	0.96
8	CONSULT GASTROENTEROLOGY	0.22	0.03	8.6	0.00
9	UPPER GI ENDOSCOPY	0.21	0.08	2.8	0.00
10	ISTAT, VBG AND LACTATE	0.21	0.19	1.1	0.47
11	Oral Electrolyte Solution (Bowel Prep)	0.17	0.03	5.3	0.00
12	OCTREOTIDE INFUSION	0.17	0.01	11.7	0.00
13	TRANSFUSE FFP	0.16	0.16	1.0	0.91
14	Benzocaine+Tetracaine (Topical)	0.09	0.04	2.0	0.00
15	H. PYLORI AG, STOOL	0.08	0.02	4.9	0.00

Table 3 – Example orders recommended when query by admitting diagnosis of GI Hemorrhage, ranked by conditionalFreq(B|A)$_{day}$ and filtering out those with freqRatio(B|A)$_{day}$ <1. Example interpretation: Given a GI Hemorrhage, 75% of patients receive IV Pantoprazole (standard initial treatment for an acute GI bleed) within 24 hours. This is somewhat more likely (freqRatio 1.8) than for all patients in general, though even the baseline of 42% is relatively common as IV Pantoprazole is used for non-GI bleed scenarios (e.g., prophylaxis against stress ulcers). For comparison, the Pantoprazole IV continuous infusion is less common (51%), but has a higher relative likelihood (freqRatio 16.0), as it is used almost exclusively in the treatment of GI bleeds.

Ranking Method	**Time Span**	**Ratio Filter**	**Recall**	**Precision**	**F1-Score**	**Weighted Recall**	**Weighted Precision**	**Weighted F1-Score**
Random			1%	2%	1%	1%	1%	1%
Baseline Freq*			17%	26%	19%	3%	24%	4%
Conditional Freq	Any		22%	31%	23%	5%	29%	6%
Conditional Freq	Hour		19%	27%*	20%	5%	17%	6%
Conditional Freq	Day		**27%**	**37%**	**28%**	7%	**37%**	9%
Conditional Freq	Day	Yes	9%	17%	11%	15%	14%	**12%**
Freq Ratio	Day		8%	12%	9%	**17%**	8%	10%

Table 4 – Average accuracy statistics for recommendation methods across 1,903 test patients comparing 10 system recommended orders vs. actual orders occurring within 24 hours. The Conditional Freq ranked methods are subdivided by what time span *t* that their item association counting accepts. The last pair of methods use the Freq Ratio for filtering (excluding recommendations with Freq Ratio <1) or ranking. Bolded entries represent the best value for each metric.

*All metrics are compared against the Baseline Freq method as a benchmark, with all yielding $p<0.01$, except precision of the Conditional Freq (1 Hour) method, having $p = 0.08$.

Rank	Description	Frequency / Likelihood Conditional	Baseline	Ratio	p
1	COMFORT CARE MEASURES	0.11	0.02	5.22	0.00
2	LIBERALIZE VISITATION POLICY	0.08	0.02	5.09	0.00
3	LACTIC ACID (High)	0.46	0.11	4.12	0.00
4	NOREPINEPHRINE IV INFUSION	0.15	0.04	3.84	0.00
5	CALCIUM CHLORIDE IV INFUSION	0.06	0.01	3.77	0.00
6	Citrate + Sodium Bicarbonate (CRRT)	0.05	0.01	3.68	0.00
7	CONSULT TO PALLIATIVE CARE	0.15	0.04	3.60	0.00
8	OSMOLALITY, SERUM (High)	0.07	0.02	3.55	0.00
9	pH Venous (Low)	0.23	0.06	3.51	0.00
10	LUNG PROTECTIVE VENTILATION	0.07	0.02	3.49	0.00

Table 5 – Inverted query example showing the top "recommendations" for items that occur *prior* to a query item of patient death, ranked by $\text{freqRatio}(B|A)_{\text{week}}$. This recognizes that many deaths are anticipated with a greater likelihood for ordering "Comfort Care Measures" and "Liberalize Visitation Policy," representing reprioritization of care for patients with expected imminent death. Complementary to that are deaths preceded by aggressive life-supporting ICU interventions including vasopressors (norepinephrine), continuous renal replacement therapy (CRRT), and mechanical ventilation for ARDS (lung protective ventilation protocol). Inverse queries can appropriately "recommend" non-order items such as abnormal lab values as well, in this case recognizing that lactic acidosis (high lactic acid) and acidemia (low pH) disproportionately precede death.

	Death	Any ICU
Evaluation period	30 days	1 week
Patients screened	1,905	1,905
Patients evaluated, excluding those with outcome occurring during 24 hour query period	1,898	1,765
Patients with outcome subsequently occurring during evaluation period	44 (2.3%)	55 (3.1%)
ROC AUC score for association prediction	0.88	0.78

Table 6 – ROC area-under-curve prediction metrics for 30 day mortality and 1 week requirement for ICU intervention (ventilator, vasopressor infusion, CRRT) based upon 1,905 test patients' first 24 hours of query clinical items.

Discussion

The item association system developed above, analogous to commercial recommender systems, recommends physician orders and predicts clinical outcomes based on statistics data-mined from electronic medical records. As illustrated in Table 4, personalizing order recommendations with the Conditional Freq ranking method improves accuracy compared to the standard Baseline Freq benchmark method that only functions as a general "best seller" list, recommending the overall most common orders, irrespective of query items. Demonstrated again is the importance of temporal information[11], with accuracy optimized when the association time span *t* is comparable to the evaluation time frame. Specifically, given the test evaluation period of 24 hours, the optimal association time span is one day.

Qualitative examples like Table 3 indicate that Freq Ratio based methods can provide more specifically relevant recommendations, but these approaches inherently perform worse by standard accuracy metrics, as confirmed in Table 4. While standard accuracy metrics favor common items, it is more impressive to correctly predict a rare item (e.g., pantoprazole infusion) than the relatively mundane correct prediction of a common item (e.g., Type & Screen). Alternative metrics, the inverted frequency weighted precision and recall, are introduced here to preferentially score prediction of uncommon items. Interestingly, the Conditional Freq method that performs best on standard accuracy metrics still performs best by the weighted precision metric. It is only for weighted recall that the Freq Ratio based methods show improvement (3% to 17%, $p<0.01$). This reinforces the notion that the two approaches serve different purposes and can both be useful depending on the goals of the query.

Table 6 reports the association framework's ability to predict clinical outcomes with ROC AUC of 0.88 for 30 day mortality and 0.78 for requiring ICU intervention within 1 week of hospitalization. These are comparable to state-of-the art prognosis scoring systems such as APACHE, MPM, and SAPS with scores ranging from 0.75 to 0.90 for predicting hospital mortality[17] and CURB-65, PSI, SCAP, and REA-ICU with scores ranging from 0.69 to 0.81 for predicting early ICU admission[18]. Other prediction possibilities could include hospital length of stay, readmissions, and many others, though the virtue of the framework is that it can predict any item labeled as an outcome event with minimal incremental effort, opening the tempting possibility of generating order recommendations based on predicted outcomes.

In closing, this work takes another step towards mature clinical decision support systems to unlock the Big Data potential of electronic medical records by enhancing a clinical order recommendation framework with

additional non-order data to better define clinical contexts, reporting of significance statistics for individual recommendations to further aid interpretability, demonstrating multiple evaluation metrics to discern common from specifically relevant items, and extending the application towards predicting clinical outcomes.

Acknowledgements

Project supported by the Stanford Translational Research and Applied Medicine (TRAM) program in the Department of Medicine (DOM). 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 Tanya Podchiyska of the STRIDE (Stanford Translational Research Integrated Database Environment) 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. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Services, H. 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. *Federal register* **77,** 54163–292 (2012).
2. Kaushal, R., Shojania, K. G. & Bates, D. W. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. *Archives of Internal Medicine* **163,** 1409–1416 (2003).
3. Overhage, J. & Tierney, W. A randomized trial of "corollary orders" to prevent errors of omission. *Journal of the American Medical Informatics Association* **4,** 364–75 (1997).
4. Bates, D. & Kuperman, G. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *Journal of the American Medical Informatics Association* **10,** 523–530 (2003).
5. De Lissovoy, G. Big data meets the electronic medical record: a commentary on "identifying patients at increased risk for unplanned readmission". *Medical care* **51,** 759–60 (2013).
6. Moore, K. D., Eyestone, K. & Coddington, D. C. The big deal about big data. *Healthcare financial management : journal of the Healthcare Financial Management Association* **67,** 60–6, 68 (2013).
7. Doddi, S., Marathe, a, Ravi, S. S. & Torney, D. C. Discovery of association rules in medical data. *Medical informatics and the Internet in medicine* **26,** 25–33 (2001).
8. Klann, J., Schadow, G. & Downs, S. M. A method to compute treatment suggestions from local order entry data. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* **2010,** 387–91 (2010).
9. Klann, J., Schadow, G. & McCoy, J. M. A recommendation algorithm for automating corollary order generation. *AMIA Annual Symposium Proceedings* **2009,** 333–7 (2009).
10. Wright, A. & Sittig, D. F. Automated development of order sets and corollary orders by data mining in an ambulatory computerized physician order entry system. *AMIA Annual Symposium Proceedings* **2006,** 819–823 (2006).
11. Chen, J. R. A. Mining for Clinical Expertise in (Undocumented) Order Sets to Power an Order Suggestion System. *Proceedings of the 2013 AMIA Summit on Clinical Research Informatics* (2013).
12. Linden, G., Smith, B. & York, J. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing* **7,** 76–80 (2003).
13. Lowe, H. J., Ferris, T. a, Hernandez, P. M. & Weber, S. C. STRIDE--An integrated standards-based translational research informatics platform. *AMIA Annual Symposium Proceedings* **2009,** 391–5 (2009).
14. Wright, A. & Bates, D. W. Distribution of Problems, Medications and Lab Results in Electronic Health Records: The Pareto Principle at Work. *Applied clinical informatics* **1,** 32–37 (2010).
15. Manning, C. D. & Schütze, H. *Foundations of Statistical Natural Language Processing*. *Computational Linguistics* **26,** 277–279 (MIT Press, 1999).
16. Shani, G. & Gunawardana, A. Evaluating Recommendation Systems. *Recommender Systems Handbook* **12,** 1–41 (2011).
17. Lemeshow, S. & Le Gall, J. R. Modeling the severity of illness of ICU patients. A systems update. *JAMA : the journal of the American Medical Association* **272,** 1049–55 (1994).
18. Renaud, B. *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. *Critical care (London, England)* **13,** R54 (2009).