Data-Mining Electronic Medical Records for Clinical Order Recommendations: Wisdom of the Crowd or Tyranny of the Mob?

Jonathan H. Chen, MD, PhD[1,2] and Russ B. Altman, MD, PhD[3,4*]
1 Center for Innovation to Implementation (Ci2i), Veterans Affairs Palo Alto Health Care System, Palo Alto, CA,
2 Center for Primary Care and Outcomes Research (PCOR), Stanford University, Stanford, CA
3 Department of Medicine, Stanford University, Stanford, CA
4 Departments of Bioengineering and Genetics, Stanford University, Stanford, CA
*To whom correspondence should be addressed. E-mail: russ.altman@stanford.edu

Abstract

Uncertainty and variability is pervasive in medical decision making with insufficient evidence-based medicine and inconsistent implementation where established knowledge exists. Clinical decision support constructs like order sets help distribute expertise, but are constrained by knowledge-based development. We produced a data-driven order recommender system that mines expertise from structured electronic medical record data on >19K hospital patients to predict clinical orders and outcomes. We present the first structured validation of automatically developed clinical decision support content against an objective external standard by identifying orders referenced in clinical practice guidelines. For example scenarios of chest pain, gastrointestinal hemorrhage, and pneumonia in hospital patients, the automated method identifies guideline referenced orders with ROC AUCs (c-statistics) (0.89, 0.95, 0.83) that improve upon statistical prevalence benchmarks (0.76, 0.74, 0.73) and pre-existing human-expert authored order sets (0.81, 0.77, 0.73), offering confidence when extrapolating to more typical clinical scenarios where well-defined external standards do not exist.

Introduction

Medical decision making is fraught with uncertainty, reflected in a third of surgeries to place pacemakers or ear tubes[1] and over forty percent of recommendations for the management of cardiac disease[2] without adequate evidence to confirm or deny their efficacy. Evidence-based medicine seeks to fill these gaps, but even with disruptive reforms[3], the expanding breadth and evolving complexity of medical practice ensures that high-quality prospective data will perpetually lag behind the need to answer clinical questions. Even when high-quality evidence is available, inconsistency in distribution and implementation can result in wide practice variability, such as a quarter of patients with a heart attack not receiving aspirin[4].

Clinical decision support (CDS) constructs such as order sets and templates reinforce consistency and compliance with best-practices by systematically distributing expertise[5,6], but their development is limited by a top-down, knowledge-based approach. This approach requires manual production and maintenance that is only feasible for a limited number of common scenarios[7]. The progressive adoption of electronic medical records (EMR) creates the opportunity for a Big Data[8] approach of crowd-sourcing clinical expertise from the bottom-up, tapping into the collective experience of many practitioners by automatically generating CDS content in a learning health system[9–11].

Background

Prior work in automated CDS content development includes reports of association rules and Bayesian networks relating orders and diagnoses, as well as unsupervised clustering of clinical orders[12–15]. We developed an item-based association framework to generate clinical order recommendations[16] analogous to Netflix or Amazon.com's "Customer's who bought A also bought B" system[17], extracting the expertise hidden in the patterns of clinical orders (e.g., labs, medications, imaging) that concretely manifest clinical decision making. This recommender predicts real clinical orders (improving precision at ten recommendations from a baseline of 26% to 37%) as well as clinical outcomes such as mortality and intensive care unit interventions (c-statistics of 0.88 and 0.78, respectively), comparable with state-of-the-art prognosis scoring systems[18].

Despite these encouraging results, a recurring concern for automatically generated CDS content is that *common* clinical decisions derived from the wisdom of the crowd[19] do not entail "good" or appropriate decisions. Validating the quality of recommender systems is challenging as there is not a well-defined or generally accepted definition of a "good" recommendation[20]. We previously demonstrated an internal validation by predicting clinician behavior and clinical outcomes[18], while others have assessed coverage of manually authored order sets[15] and qualitative assessment by a couple clinician reviewers[12,13]. Limited evaluations exist against objective external standards. Here we contribute a structured evaluation of automated order recommendations against the external standard of clinical practice guidelines, representing the appropriate standard of care for sample scenarios including chest pain, gastrointestinal hemorrhage (GI bleed), and pneumonia in hospital patients.

Methods

As described previously[18], we extracted patient data deidentified of protected health information for inpatient hospitalizations at Stanford University Hospital in 2011 from the STRIDE clinical data warehouse[21]. The structured data covers patient encounters from their initial (emergency room) presentation until hospital discharge, including >19K distinct patients with >5.4M instances of >17K distinct clinical items. Clinical items include medication, laboratory, imaging, and nursing orders, as well as non-order items for lab results, ICD9 diagnosis codes, and patient demographics. Applying the "80/20 rule" in the form of a power law distribution[22], rarely used clinical items were removed from consideration to reduce the effective item count from >17K to the top 1.5K (9%), while still covering 5.1M (94%) of the item instances. Furthermore, infrastructure orders that are commonly part of treatment processes or admission protocols (e.g., vital signs, notify MD, regular diet, patient transport, and all other nursing and PRN medication orders) were excluded as they rarely reflect meaningful clinical decisions. These exclusions left 811 clinical order items as candidates for recommendation.

To develop an external reference standard for order quality, we evaluated clinical practice guidelines from the National Guideline Clearinghouse (http://www.guideline.gov) that inform the management of selected hospital ICD9 admission code groups for chest pain[23,24], GI Bleed[25–28], and pneumonia[29,30]. Diagnoses were selected based on the existence of relevant guidelines and a significant quantity of clinical data examples. The 811 candidate clinical orders were labeled based on whether a guideline explicitly mentioned them as appropriate to consider (e.g., treating pneumonia with levofloxacin), or implied them (e.g., bowel preps and NPO diet orders are implicitly necessary to fulfill explicitly recommended endoscopy procedures for GI Bleeds). Given the non-specific nature of admission diagnoses, separate guideline recommendations for management of ulcerative, variceal, and lower GI bleeding were included as the specific etiology is generally unknown at the time of hospital admission. Similarly, while we included chest pain guideline recommendations focused on empiric management and diagnosis, we excluded recommendations prescribing treatment for confirmed diagnoses (e.g., clopidogrel for heart attack, NSAIDs and colchicine for pericarditis) as such cases presumably would have been admitted under the specific diagnosis, rather than the undifferentiated "chest pain" syndrome.

Automated order recommendations were generated from our previously described methods[16]. Specifically, based on Amazon's product recommender[17], an intensive pre-computation step collects frequency statistics for all clinical item instances and co-occurrences on a randomly selected training set of 15,629 patients to build an item association matrix, based on the definitions in Table 1. Counting by patients affords a natural interpretation of 2x2 contingency tables for pairs of items, from which various association statistics can be derived (e.g., odds ratio (OR), relative risk (RR), positive predictive value (PPV), sensitivity, baseline prevalence, Fisher's P-value)[31].

Notation	Definition
$\mathbf{n_A}$	Number of patients where item A occurs
$\mathbf{n_{ABt}}$	Number of patients where item B follows item A within time t
N	Total number of patients

Table 1 - Pre-computed frequency statistics for clinical items. Ignoring repeats generates patient counts where items occur.

The order recommender used admission diagnoses as query items A to generate lists of all candidate clinical order items B occurring within the order-dense first 24 hours of each admission, score-ranked by one of the association statistics. These score-ranked order item lists were evaluated against the guideline order reference set by receiver operating characteristic (ROC) analysis and recommendation accuracy in terms of precision (PPV) and recall (sensitivity) when considering only the top *K* recommendations.

We selected pre-authored order sets available in the hospital EMR that were relevant for each admission diagnosis to provide an order selection performance benchmark. The 811 candidate clinical order items were labeled based on inclusion in a relevant order set. Order set items were organized by implied priority based on whether they were selected by default (<5%), directly available but not default-selected, or only accessible under sub-menus (~20%), the last of which were not counted when labeling order set item inclusion.

A limitation of conventional order set design is that they generally present all n_O possible items (up to 102 in the case of chest pain) without a ranking method to convey the relative importance of items (default selections and submenus differentiated <25% of order set items). This required some approximation when using the human-expert authored order sets as a benchmark method for identifying guideline reference orders as they effectively yield only two possible score-ranks for candidate items resulting in a single discrete point on the ROC "curve" and the accuracy vs. top *K* items plots using an arbitrary ranking of items within order sets.

Results

Table 2 reports summary statistics on patient information available, guideline reference orders, and pre-authored order set items for each of the admission diagnoses considered. Table 3 contains recommendation examples for the chest pain admission diagnosis with association statistics and reference labels. Figure 1 depicts ROC curves assessing order recommendation discrimination of guideline reference orders. Figure 2 depicts recommendation accuracy for increasing number of *K* items considered, illustrating the tradeoff between precision and recall, and performance for more practical small values of *K*.

Admission Diagnosis (ICD9)	Training Patients	Guideline Orders Referenced	Order Set Orders	Guideline Orders in Order Sets
GI Bleed (578)	282	38	51	22
Chest Pain (786.5)	433	32	102	23
Pneumonia (486)	206	51	42	25

Table 2 – Admission diagnoses evaluated, number of patients in training dataset, orders referenced in clinical practice guidelines, orders available in pre-authored order sets, and the intersection between the two.

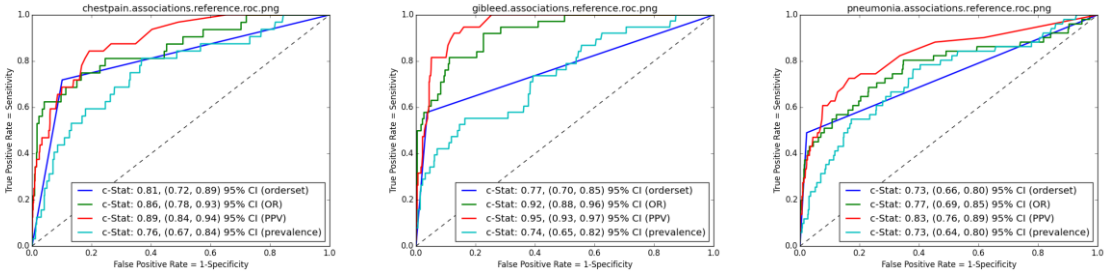

Figure 1 – Receiver operating characteristic (ROC) curves for predicting clinical practice guideline reference orders based on automated recommender methods using different score-ranking options (PPV, OR, baseline prevalence, and presence in pre-authored order sets). Area-under-curve (AUC) reported as c-statistics with 95% confidence intervals empirically estimated by sampling items with replacement 1000 times.

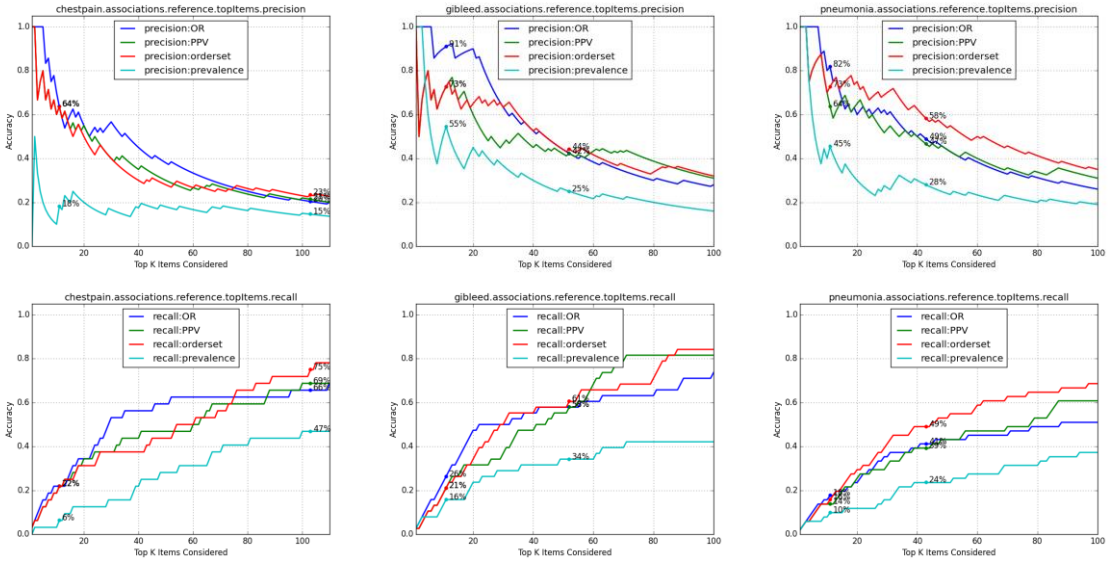

Figure 2 – Recommender accuracy (precision or recall) for predicting guideline reference orders as a function of the number of top *K* recommendations considered (up to 100) when sorting by different score-ranking options (OR, PPV, prevalence, and presence in pre-authored order sets). Data labels added for $K = 10$ and n_O, where n_O = Number of items available in the respective order sets.

Item Description	Prevalence	PPV	OR	P-Fisher	Order Set/ Guideline
POC ISTAT TROPONIN I	16.3%	71.4%	14.4	1.6E-148	1 / 1
ECG 12-LEAD	51.8%	92.8%	12.7	8.0E-80	1 / 1
Nitroglycerin (Sublingual)	1.1%	9.2%	11.2	2.3E-25	0 / 1
CONSULT TO CARDIOLOGY	4.6%	28.4%	9.6	8.2E-64	1 / 1
D - DIMER (ELISA)	1.4%	9.7%	9.4	5.2E-24	1 / 0
Aspirin (Oral)	24.4%	68.4%	7.2	2.4E-85	1 / 1
CK, MB (MASS)	15.7%	51.3%	6.1	1.5E-68	1 / 0
TROPONIN I	23.8%	62.4%	5.6	3.3E-67	1 / 1
Clopidogrel (Oral)	5.6%	20.6%	4.7	1.5E-27	1 / 0
CAR CATH CORS POSSIBLE	2.6%	9.9%	4.4	6.7E-14	1 / 1
HEPARIN ACTIVITY LEVEL	6.1%	18.9%	3.8	1.7E-20	0 / 0
LIPID PANEL WITH DIRECT LDL	8.5%	24.5%	3.7	4.6E-24	1 / 0
NT - PROBNP	10.0%	26.3%	3.4	6.4E-23	1 / 1
Nitroglycerin (Topical)	2.2%	6.2%	3.2	8.1E-07	1 / 1
ECHO - DOBUTAMINE STRESS ECHO	1.2%	3.2%	3.0	6.2E-04	1 / 1

Table 3 – Example top order recommendations occurring within 24 hours of admission diagnosis of Chest Pain (ICD9: 786.5), sorted by odds ratio (OR). Additional metrics include prevalence (pre-test probability), positive predictive value (post-test probability), and P-value by Fisher's exact test. Binary labels are assigned if the order exists in pre-authored order sets or clinical practice guidelines.

Discussion

Figure 1 illustrates c-statistics (AUC) when using automated recommendations to identify orders referenced in clinical practice guidelines, improving upon the benchmarks set by pre-authored order sets and baseline prevalence (pre-test probability). Figure 2 illustrates a similar trend for recommendation precision at small values of *K* recommendations considered, which is more relevant to a satisfactory end-user experience. These results support the hypothesis that automated methods for decision support content development will yield appropriate recommendations consistent with standards of care. However, the validation is only demonstrated for this sample selection of admission diagnosis scenarios. Given that the pre-authored order sets were available to clinicians during the training data period, the direct incremental value of the example recommendations presented is limited, especially if they tend to reproduce existing order sets. The more important value of this work comes from extrapolating to more typical clinical scenarios that are too specific or complex, such that clinical practice guidelines and pre-authored order sets do not apply or perhaps do not even exist (e.g., management of an admission diagnosis of "altered mental status" (ICD9: 780.97), or a patient with a combination of medications orders for furosemide, spironolactone, and lactulose). In such cases where reference standards for high-quality orders do not exist, automated learning methods still provide data-driven recommendations based on other practitioners' experience, with the example cases analyzed here providing confidence in the quality of those automated recommendations.

Recommendation quality also depends on the complementary goal of *not* recommending "inappropriate" orders. For our example cases, guidelines recommend against the routine use of IV hydrocortisone (stress does steroids) and filgrastim (granulocyte colony stimulating factor) in pneumonia and Factor VIIa in GI bleeding. The automated order recommendations appropriately score these orders with low odds ratios <1 and PPVs <3%. CK-MB for chest pain is the notable exception where automated recommendations endorse an order which guidelines explicitly reference as inappropriate. CK-MB is a cardiac biomarker for heart attacks largely made obsolete by more accurate troponin testing, but the practice patterns at this hospital (revealed by the recommender statistics) demonstrate the routine habit of ordering CK-MB (likely exacerbated by the inclusion of CK-MB in the pre-authored chest pain order set!), perpetuating a debate as to whether CK-MB tests still provide any value[32].

The primary limitation of this study approach is the inherent complexity of medicine that defies the existence of a gold standard for general clinical decision-making quality given patient-to-patient variability and insufficient prospective evidence. Clinical practice guidelines provide a reference standard, yet they often offer deliberately vague, conditional, and sometimes conflicting recommendations. For example, multiple guidelines recommended providing appropriate counseling for patients and families and performing hemodynamic stability assessment with resuscitative stabilization, yet the former is not mappable to concrete actions and how to fulfill the latter is left for the clinician to interpret. Some guideline recommended tests such as Streptococcus antigen testing for pneumonia and treatments like terlipressin for esophageal variceal bleeding are not routinely available in the hospital evaluated (the latter is not even FDA approved in the US, with the guideline produced in the UK). One chest pain guideline specifically recommended against the use of stress EKG testing and natriuretic peptides in the diagnostic workup for hospitalized patients with chest pain, while the other specifically referenced both as reasonable options.

The complexity and uncertainty of medical-decision making requires clinicians to accumulate expert knowledge through extensive experiential learning, yet they rarely document their expertise in any formal manner. Information retrieval methods will enable the systematic extraction and dissemination of this undocumented collective wisdom by translating the end clinical data into a reproducible and executable form of expertise, unlocking the Big Data potential of electronic medical records.

Acknowledgements

Project supported by the Stanford Translational Research and Applied Medicine (TRAM) program in the Department of Medicine (DOM). 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 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.

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

References

1. Committee on Quality of Health Care in America I of M. Crossing the Quality Chasm A New Health System for the 21st Century. 2001.
2. Allen JM, Kramer JM, Califf RM, Jr SCS. Scientific Evidence Underlying the ACC / AHA. JAMA. 2009;301(8):831–41.
3. Lauer MS, Bonds D. Eliminating the "expensive" adjective for clinical trials. Am Heart J. Elsevier B.V.; 2014 Apr;167(4):419–20.
4. Committee on Quality of Health Care in America I of M. Crossing the quality chasm: a new health system for the 21st century. JAMA: The Journal of the American Medical …. 2001.
5. Kaushal R, Shojania KG, Bates DW. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. Shojania KG, Duncan BW, McDonald KM, Wachter RM, Markowitz AJ, editors. Arch Intern Med. Am Med Assoc; 2003;163(12):1409–16.
6. Overhage J, Tierney W. A randomized trial of "corollary orders" to prevent errors of omission. J Am Med Informatics Assoc. 1997;4(5):364–75.
7. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc. 2003;10(6):523–30.
8. Moore KD, Eyestone K, Coddington DC. The big deal about big data. Healthc Financ Manage. 2013 Aug;67(8):60–6, 68.
9. Longhurst C a., Harrington R a., Shah NH. A "Green Button" For Using Aggregate Patient Data At The Point Of Care. Health Aff. 2014 Jul 8;33(7):1229–35.
10. Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. N Engl J Med. 2011 Nov 10;365(19):1758–9.
11. Krumholz HM. Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. Health Aff. 2014 Jul 8;33(7):1163–70.
12. Klann J, Schadow G, Downs SM. A method to compute treatment suggestions from local order entry data. AMIA Annu Symp Proc. 2010 Jan;2010:387–91.
13. Klann J, Schadow G, McCoy JM. A recommendation algorithm for automating corollary order generation. AMIA Annu Symp Proc. 2009 Jan;2009:333–7.
14. 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. American Medical Informatics Association; 2006;2006:819–23.
15. Zhang Y, Padman R, Levin JE. Paving the COWpath: data-driven design of pediatric order sets. J Am Med Inform Assoc. 2014 Apr 1;
16. Chen JH, Altman RB. Mining for clinical expertise in (undocumented) order sets to power an order suggestion system. AMIA Jt Summits Transl Sci Proc AMIA Summit Transl Sci. 2013 Jan;2013:34–8.
17. Linden G, Smith B, York J. Amazon.com recommendations: item-to-item collaborative filtering. Smith B, editor. IEEE Internet Comput. IEEE; 2003;7(1):76–80.
18. Chen JH, Altman RB. Automated Physician Order Recommendations and Outcome Predictions by Data-Mining Electronic Medical Records. AMIA Jt Summits Transl Sci Proc AMIA Summit Transl Sci. 2014;
19. Surowiecki J. The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.
20. Shani G, Gunawardana A. Evaluating Recommendation Systems. Ricci F, Rokach L, Shapira B, Kantor PB, editors. Recomm Syst Handb. Springer; 2011;12(19):1–41.
21. Lowe HJ, Ferris T a, Hernandez PM, Weber SC. STRIDE--An integrated standards-based translational research informatics platform. AMIA Annu Symp Proc. 2009 Jan;2009:391–5.
22. Wright A, Bates DW. Distribution of Problems, Medications and Lab Results in Electronic Health Records: The Pareto Principle at Work. Appl Clin Inform. 2010 Jan;1(1):32–7.
23. Practice C. Chest pain of recent onset pain or discomfort of suspected cardiac origin. 2010;(March).
24. Guideline HC. Diagnosis and Treatment of Chest Pain and Acute Coronary Syndrome (ACS) Health Care Guideline and Order Set : Diagnosis and Treatment of Chest Pain and Acute Coronary Syndrome (ACS). 2012;
25. Laine L, Jensen DM. Management of patients with ulcer bleeding. Am J Gastroenterol. Nature Publishing Group; 2012 Mar;107(3):345–60; quiz 361.
26. Criteria ACRA. Radiologic Management of Lower Gastrointestinal Bleeding. 2011;8–13.
27. Evidence NHS, Practice C. Acute upper gastrointestinal bleeding : management. 2012;(April 2007).
28. Gastroenterology W, Global O. Esophageal varices. 2014;(January).
29. Society BT. Guidelines for the Management of Community Acquired Pneumonia in Adults Update 2009 A Quick Reference Guide British Thoracic Society. 2009;(November).
30. Mandell L a, Wunderink RG, Anzueto A, Bartlett JG, Campbell GD, Dean NC, et al. Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. Clin Infect Dis. 2007 Mar 1;44 Suppl 2(Suppl 2):S27–72.
31. Finlayson SG, LePendu P, Shah NH. Building the graph of medicine from millions of clinical narratives. Sci Data. 2014 Sep;1:140032.
32. Collinson PO, van Dieijen-Visser MP, Pulkki K, Hammerer-Lercher A, Suvisaari J, Ravkilde J, et al. Evidence-based laboratory medicine: how well do laboratories follow recommendations and guidelines? The Cardiac Marker Guideline Uptake in Europe (CARMAGUE) study. Clin Chem. 2012 Jan;58(1):305–6.