# Inpatient Clinical Order Patterns Machine-Learned From Teaching Versus Attending-Only Medical Services

**Jason K. Wang,[1] Alejandro Schuler,[2] Nigam H. Shah, MBBS, PhD,[2] Michael T. M. Baiocchi[3], PhD, Jonathan H. Chen, MD, PhD[4]**

[1]Mathematical & Computational Science Program, Stanford University, Stanford, CA, USA;
[2]Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA;
[3]Prevention Research Center, Stanford University, Stanford, CA, USA;
[4]Department of Medicine, Stanford University, Stanford, CA, USA

## Abstract

*Clinical order patterns derived from data-mining electronic health records can be a valuable source of decision support content. However, the quality of crowdsourcing such patterns may be suspect depending on the population learned from. For example, it is unclear whether learning inpatient practice patterns from a university teaching service, characterized by physician-trainee teams with an emphasis on medical education, will be of variable quality versus an attending-only medical service that focuses strictly on clinical care. Machine learning clinical order patterns by association rule episode mining from teaching versus attending-only inpatient medical services illustrated some practice variability, but converged towards similar top results in either case. We further validated the automatically generated content by confirming alignment with external reference standards extracted from clinical practice guidelines.*

## Introduction

Healthcare in the US often falls short of optimal, evidence-based care, with overall compliance with evidence-based guidelines ranging from 20-80%.[1] Even with recent reforms, evidence-based medicine from randomized controlled trials cannot keep pace with the ever-growing breadth of clinical questions, with only 11% of guideline recommendations supported by high-quality evidence.[2] This variability and uncertainty in medical practice is further exacerbated by a medical knowledge base that is perpetually expanding beyond the cognitive capacity of any individual.[3] A clinician is thus left to synthesize vast streams of information in an attempt to make the best decision for each individual patient. As such, medical practice routinely relies on anecdotal experience and individual expert opinion.

To address these issues, healthcare organizations have increasingly adopted clinical decision support (CDS) systems. CDS aims to reinforce best-practices by distributing knowledge-based content through order sets, templates, alerts, and prognosis scoring systems.[4] Here we focus specifically on clinical orders (e.g. lab tests, medications, imaging exams) as concrete manifestations of point-of-care decision making. Computerized provider order entry (CPOE) typically occurs on an "a la carte" basis, where clinicians search for and select orders to trigger subsequent clinical actions (e.g. pharmacy dispensing and nurse administration of a medication, laboratory analysis of blood tests, consultation to a specialist). Because clinician memory and intuition can be error-prone, health system committees manually curate template order sets to distribute standard practice guidelines specific to a diagnosis or medical procedure.[5] This top-down approach enables clinicians to draw clinical orders from pre-constructed, human-authored templates when treating common scenarios (e.g. pneumonia, stroke).

Existing approaches to CDS increase consistency and compliance with best practices,[6-7] but production of this content is limited in scale by a committee-driven, manual production process that oftentimes requires the collaboration of a multi-disciplinary team of physicians, nurses, and department heads. Once an order set is published and made available to clinicians, ongoing maintenance is required to keep it up to date with new evidence, technology, epidemiology, and culture.[8] As such, one of the "grand challenges" in CDS is to automatically generate content by data-mining clinical data sources from the bottom-up.[9] In the era of electronic health record (EHR) data, there is an opportunity to create a data-driven CDS system that leverages the aggregate expertise of many healthcare providers and automatically adapts to the ongoing stream of practice data.[10] This would fulfill the vision of a health system that continuously learns from real-world data and translates them into usable, point-of-care information for clinicians. Prior research into data-mining for decision support content includes association rules, Bayesian networks, and unsupervised clustering of clinical orders and diagnoses.[11-18] In our prior work, inspired by similar information retrieval problems, we developed a data-driven clinical order recommender engine[19] analogous to Netflix and Amazon.com's "customers who bought A also bought B" system.[20] Our engine dynamically generates order recommendations based on real-world clinical practice patterns represented in EHR data.

In psychology, the "wisdom of the crowd" phenomenon purports that the collective assessment of a group of non-expert individuals is generally as good as, if not better than, that of individual experts.[21] In the context of data-driven CDS for medical decision making, this translates to training machine-learning models on all available data, including patterns generated by clinicians of all levels of experience, rather than just a subset of data generated by the most experienced few. However, effective medical decision making and patient outcomes could conceivably be compromised if patterns are learned from less experienced fellows, residents, and medical students administering care. In fact, patient outcomes tend to worsen during end-of-year changeover when new, less experienced trainees enter the hospital.[22] This phenomenon gives rise to the concern over learning indiscriminately from the "wisdom of the crowd" when the crowd consists of both experienced attending physicians and teaching services that include trainees. Prior studies have examined the outcomes of specific procedures when trainees of differing levels of experience are involved in the patient care process, assessing metrics including readmission rates, mortality, procedural time, etc.[23-24] As we seek to evaluate emerging data-driven CDS systems, an essential step is understanding the implications of learning from clinical practices of varying experience and potentially of variable expertise and quality.

In this study, we investigate how clinical orders learned from two distinct clinical settings, a university teaching versus an attending-only service, influence the alignment of clinical order patterns learned from data-mining EHR data to clinical practice guidelines.

## Methods

We extracted deidentified, structured patient data from the (Epic) electronic medical record for inpatient hospitalizations in 2013 at Stanford University Medical Center via the STRIDE clinical data warehouse.[25] The dataset covers patient encounters from their initial (emergency room) presentation until hospital discharge, comprising >20,000 patients and >6.7 million instances of >23,000 distinct clinical data items.

## Data Preparation

The large majority of clinical items are medications, laboratory tests, imaging tests, and nursing orders, while non-order items include lab results, problem list entries, admission diagnosis ICD9 codes, and patient demographics on age, gender, and date of death. Medication data was normalized with RxNorm mappings[26] down to active ingredients and routes of administration. Numerical lab results were binned into categories based on "abnormal" flags established by the clinical laboratory, or being outside two standard deviations from the population mean. ICD9 codes were aggregated up to the three digit hierarchy to compress the sparsity of diagnosis categories; original detailed codes were retained if the diagnosis was sufficiently prevalent to be useful. Income levels were inferred from 2013 US census data by cross-referencing each patient's zip code with the median household income in that region. These pre-processing steps enable us to model each patient as a timeline of clinical item event instances, with each instance mapping a clinical item to a patient at a discrete time point.

With the clinical item instances following the 80/20 rule of a power law distribution, the majority of item types may be ignored with minimal information loss.[27] In this case, ignoring rare clinical items with <64 occurrences (~0.004% all instances) reduces the distinct item count to ~12% of the original, while still capturing ~95% of the individual item instances. Removing rare items further avoids spurious results with insufficient data for reliable statistical inference. Common process orders (e.g., vital signs, notify MD, regular diet, transport patient, as well as most nursing and all PRN medications) were excluded, leaving just over 1,400 candidate clinical orders for consideration.

### Cohort Selection

Using patient provider information, we prepared two patient cohorts seen by distinct general medicine services: 1) a private attending-only service (n=1774) and 2) a university teaching service (n=3404). University teaching services emphasize both medical education and patient care, thus allowing trainees (e.g. medical students, residents, and fellows) to play an active role in deciding on and administering care plans alongside attendings, nurses, and physician assistants. Although trainees discuss high-level care plans with supervising physicians at least once a day, they are almost exclusively responsible for inputting clinical orders into the EHR system. In contrast, the attending-only service is run exclusively by board-certified physicians.

### Propensity Score Matching

The private attending-only and university teaching services are within the same hospital system with access to the same pool of resources. However, the private attending service exclusively accepts patients who see clinicians in a local private health plan, while the teaching service admits patients of all health plans, including a small number of uninsured. To minimize potential biases arising from this difference, we conducted propensity score matching to balance the two patient cohorts. Using demographic data (age, gender, ethnicity, income level), initial vital signs recorded before the onset of care (temperature, pulse, respiration), and existing diagnoses as covariates, we applied a logistic regression model to compute the probability $p$ of each patient's assignment to the attending-only cohort. The propensity score is then defined as the logit function $log \frac{p}{1-p}$. Caliper matching on the propensity score resulted in balanced cohorts of 1530 patients each. A caliper threshold of 0.18 was chosen based on $0.2 \times \sigma$ where $\sigma =$ the standard deviation across all propensity scores.[28]

### Association Rule Episode Mining

Using the preprocessed data from patient encounters associated with an input patient cohort, we conducted association rule episode mining for clinical item pairs to capture historical clinician behavior. Our previously described clinical order recommender algorithm[19,29] counts co-occurrences for all clinical item pairs occurring within 24 hours to build time-stratified item association matrices. For each pair of items A and B, the co-occurrence counts accumulated can be represented as $N_{AB,t}$, the number of patients for whom item B follows A within time t, as illustrated by the pseudocode below.

```
For each patient P:
  For each item A that occurs for patient P at time $t_A$:
    For each item B that occurs for patient P at time $t_B$ where $t_B$>=$t_A$:
      If (P,A,$t_A$) or (P,B,$t_B$) not previously analyzed:
        If ($t_B$-$t_A$) <= timeThreshold and (P,A,B) newly encountered:
          Increment $N_{AB,timeThreshold}$
    Record (P,A,B) as previously encountered
  Record (P,A,$t_A$) as previously analyzed
```

These counts are then used to populate 2x2 contingency tables to compute association statistics such as baseline prevalence, positive predictive value (PPV), relative risk (RR), odds ratio (OR), and P-value by chi-square test with Yates' correction for each pair of clinical items. For a given query item (e.g. admission diagnosis), we generate a list of clinical item suggestions score-ranked by a specified association statistic. Score-ranking by PPV prioritizes orders that are *likely* to occur after the query items, while score-ranking by P-value for items with odds ratio >1 prioritizes orders that are *disproportionately* associated with the query items.[29]

We trained two distinct association models either using patient encounters from the balanced attending-only or teaching service cohorts. We then generated a predicted order list ranked by PPV from each model for 6 common diagnoses: altered mental status (ICD9: 780), chest pain (ICD9: 786.5), gastrointestinal (GI) hemorrhage (ICD9: 578.9), heart failure (ICD9: 428), pneumonia (ICD9: 486), and syncope and collapse (ICD9: 780.2).

### Guideline Reference Standard

To develop an external reference standard for order quality, a board-certified internal medicine physician manually curated reference lists based on clinical practice guidelines available from the National Guideline Clearinghouse (www.guideline.gov) and PubMed that inform the inpatient management of altered mental status,[30] chest pain,[31-32] GI hemorrhage,[33-35] heart failure,[36-37] pneumonia,[38-39] and syncope and collapse.[40] The specific diagnoses were selected based on the existence of relevant guidelines and a significant quantity of clinical data examples. Candidate clinical orders were included in the reference standard based on whether a guideline text explicitly mentioned them as appropriate to consider (e.g., treating pneumonia with levofloxacin), or heavily implied them (e.g., bowel preps and NPO diet orders are implicitly necessary to fulfill explicitly recommended endoscopy procedures for GI bleeds).

### Evaluation Metrics

PPV-ranked predicted order lists were generated from both attending-only and teaching service association models. To assess the similarity between the two predicted order lists, traditional measures of list agreement like Kendall's $\tau$ -metric[41] are not ideal as they often require identically sized, finite lists, and weigh all list positions equivalently, neglecting rank-order. To compare ranked clinical order lists, we instead calculate their agreement by Rank Biased Overlap (RBO).[42] RBO computes the average fraction of top items in common between two ordered list and is characterized by a "persistence" parameter p, the probability that an observer reviewing the top k items will continue to observe the (k+1)-th items. For our calculations, we used a default implementation parameter p of 0.98. This has the effect of geometrically weighting emphasis to the top of each list. RBO values range from 0.0 (no correlation or random list order) to 1.0 (perfect agreement of list order).

To obtain a more objective measure of quality, we also compare each predicted order list against the corresponding guideline reference list using area under the receiver operating characteristic (ROC AUC = c-statistic) and precision and recall for the top K ranked items. Comparison of such metrics will determine whether attending-only and teaching service cohorts differ in their practice of guideline aligned medicine. Non-meaningful, small absolute differences may still yield "statistically significant" P values given sufficiently large data sizes; thus, we can judge clinical settings as being comparable if their c-statistics are "bioequivalent." That is, if the 95% confidence interval for each cohort's c-statistic falls within 80-125% of the other cohort's c-statistic (definition from pharmacologic bioequivalence of generic versus brand drugs).[43]

### Results

The post-matching standardized mean differences (SMD) and p-values computed using two-sample t-tests were < 0.1 and > 0.15, respectively, across all covariates, demonstrating a statistically insignificant difference between balanced attending-only and teaching service patient cohorts. Figure 1 shows the SMD of all 25 covariates before and after propensity score matching. Post-matching results indicate that matching was necessary and effective in balancing patient cohorts for a fair comparison with respect to measured covariates.

Table 1 illustrates examples of the top clinical orders predicted by our recommender engine for an admission diagnosis of pneumonia and altered mental status trained separately on each patient cohort. The predicted order lists for pneumonia corroborate on several top clinical items including intravenous administration of levofloxacin, a drug commonly used to treat bacterial infections, and blood cultures to detect pathogens in the bloodstream. Likewise, the two predicted order lists for altered mental status agree on conducting a CT scan of the patient's head as well as a screening of blood, urine, or other body samples to investigate the patient's usage of certain drugs. To assess the similarity between order lists outputted from the attending-only and teaching service association models while still capturing the score-based ranking within each list, we use RBO. RBO values of ~0.7 for all 6 common diagnosis shown in Table 2 demonstrate strong agreement between each pair of predicted order lists. Figure 2 shows ROC curves generated for the 6 admission diagnoses. Each plot reports 3 order lists compared against the guideline reference standard, corresponding to 3 curves: attending-only predicted orders, teaching service predicted orders, and the real-world pre-authored order set manually curated by Stanford University's Medical Center serving as a benchmark. Pre-authored order sets have no inherent ranking or scoring system to convey relative importance and are thus depicted as a single discrete point on the ROC curve. Area-under-curve (AUC) is reported as c-statistics with 95% confidence intervals empirically estimated by bootstrap resampling with replacement 1000 times. Figure 3 depicts recommendation accuracy for an increasing number of K items considered, illustrating the tradeoff between precision and recall.

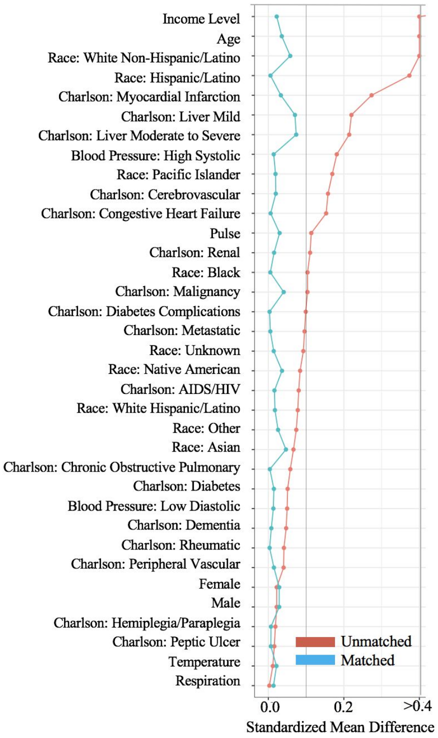


**Figure 1.** The standardized mean difference (SMD) between attending-only and teaching service cohorts across 25 covariates spanning demographic data, initial vital signs, and existing diagnoses. For a given covariate, the SMD is defined as the difference between the mean value for each cohort divided by the pooled standard deviation.

**Table 1.** Top five ranked clinical order associations for pneumonia (ICD9: 486) (top) and altered mental status (ICD9: 780) (bottom) predicted by attending-only and teaching service trained models sorted by P-value calculated by Yates' chi-squared statistic. Additional association statistics (e.g. baseline prevalence, PPV, RR) and a column denoting the presence or absence of the predicted item in the corresponding human-authored hospital order set or guideline reference standard are also included. Items with a baseline prevalence <1% are excluded to avoid statistically spurious results and ensure computationally tractable association rule episode mining. Each item represents a clinical order that a clinician can request through a CPOE system. An automated order set can be curated by selecting the top K ranked clinical orders.

| Attending-Only Service Orders | Prevalence | PPV | RR | P-Value | Order Set/ Guideline | Teaching Service Orders | Prevalence | PPV | RR | P-Value | Order Set/ Guideline |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Respiratory Isolation | 0.02 | 0.18 | 8.4 | 2x10$^{-8}$ | No/No | Azithromycin (Intravenous) | 0.10 | 0.41 | 4.7 | 4x10$^{-16}$ | Yes/Yes |
| Levofloxacin (Intravenous) | 0.14 | 0.45 | 3.4 | 4x10$^{-8}$ | Yes/Yes | Levofloxacin (Intravenous) | 0.14 | 0.46 | 3.7 | 9x10$^{-14}$ | Yes/Yes |
| Blood Culture (Aerobic & Anaerobic Bottles) | 0.50 | 0.90 | 1.9 | 5x10$^{-7}$ | Yes/Yes | Respiratory Nebulizer | 0.26 | 0.52 | 2.1 | 3x10$^{-6}$ | Yes/No |
| Blood Culture (2 Aerobic Bottles) | 0.49 | 0.88 | 1.8 | 1x10$^{-6}$ | Yes/Yes | Blood Culture (Aerobic & Anaerobic Bottles) | 0.59 | 0.87 | 1.5 | 4x10$^{-6}$ | Yes/Yes |
| Respiratory Culture | 0.11 | 0.30 | 2.9 | 3x10$^{-4}$ | Yes/Yes | Acetaminophen (Rectal) | 0.02 | 0.11 | 6.3 | 5x10$^{-6}$ | Yes/No |

| Attending-Only Service Orders | Prevalence | PPV | RR | P-Value | Order Set/ Guideline | Teaching Service Orders | Prevalence | PPV | RR | P-Value | Order Set/ Guideline |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CT Head | 0.22 | 0.62 | 2.9 | 2x10$^{-11}$ | Yes/Yes | CT Head | 0.25 | 0.63 | 2.7 | 6x10$^{-18}$ | Yes/Yes |
| Drugs of Abuse Screen Urine | 0.09 | 0.31 | 3.7 | 1x10$^{-7}$ | Yes/Yes | Ammonia Plasma | 0.08 | 0.26 | 3.9 | 1x10$^{-11}$ | Yes/Yes |
| Consult to Neurology | 0.03 | 0.15 | 5.2 | 8x10$^{-6}$ | No/Yes | Volatile Screen | 0.09 | 0.22 | 2.6 | 1x10$^{-5}$ | No/Yes |
| Acetaminophen Serum | 0.03 | 0.15 | 5.2 | 8x10$^{-6}$ | No/No | Drugs of Abuse Screen Urine | 0.17 | 0.32 | 2.1 | 3x10$^{-5}$ | No/Yes |
| Salicylate Level | 0.03 | 0.13 | 5.8 | 8x10$^{-6}$ | No/Yes | Rifaximin (Oral) | 0.03 | 0.08 | 3.7 | 1x10$^{-3}$ | No/No |

**Table 2.** Rank Biased Overlap (RBO)[42] computed between attending-only and teaching service order lists, score-ranked by PPV, predicted for 6 common diagnoses: altered mental status (ICD9: 780), chest pain (ICD9: 786.5), gastrointestinal (GI) hemorrhage (ICD9: 578.9), heart failure (ICD9: 428), pneumonia (ICD9: 486), and syncope and collapse (ICD9: 780.2). RBO computes the average fraction of top items in common, geometrically weighting all 1468 or 1474 candidate clinical order items based on a scoring metric (e.g. PPV) for the attending-only and teaching service cohorts, respectively. RBO values of ~0.7 indicate strong overlap between order lists generated by the two cohorts.

| Diagnosis | Rank Biased Overlap Attendee-Only vs. Teaching Service |
|---|---|
| Altered Mental Status (780) | 0.74 |
| Chest Pain (786.5) | 0.72 |
| Gastrointestinal Hemorrhage (578.9) | 0.72 |
| Heart Failure (428) | 0.68 |
| Pneumonia (486) | 0.72 |
| Syncope and Collapse (780.2) | 0.68 |

**Figure 2.** ROC plots for the 6 common diagnoses. Each plot compares an order set authored by the hospital and automated predictions from attending-only and teaching service association models against the guideline reference standard. In all cases excluding heart failure, both model-predicted order lists show substantially larger c-statistics than the respective order set benchmark. As the manually-curated hospital order set has no inherent ranking, it is plotted as a single point in which all order set items are considered.

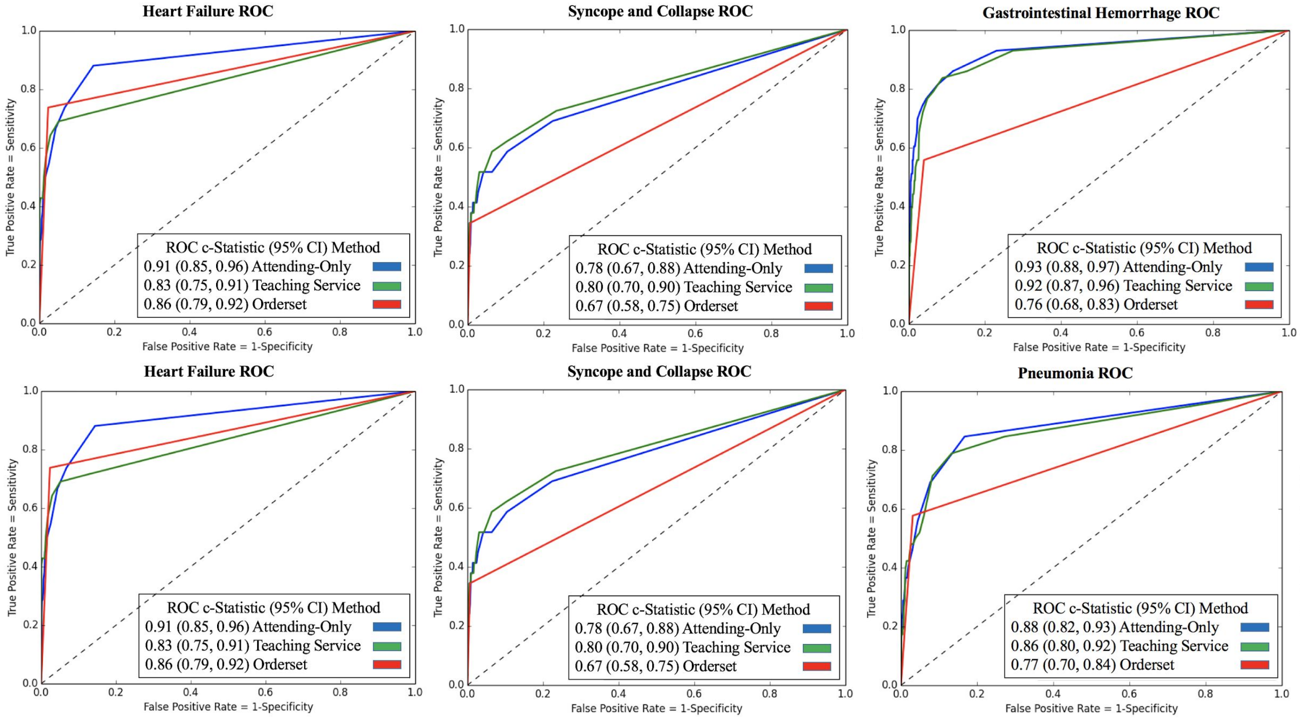
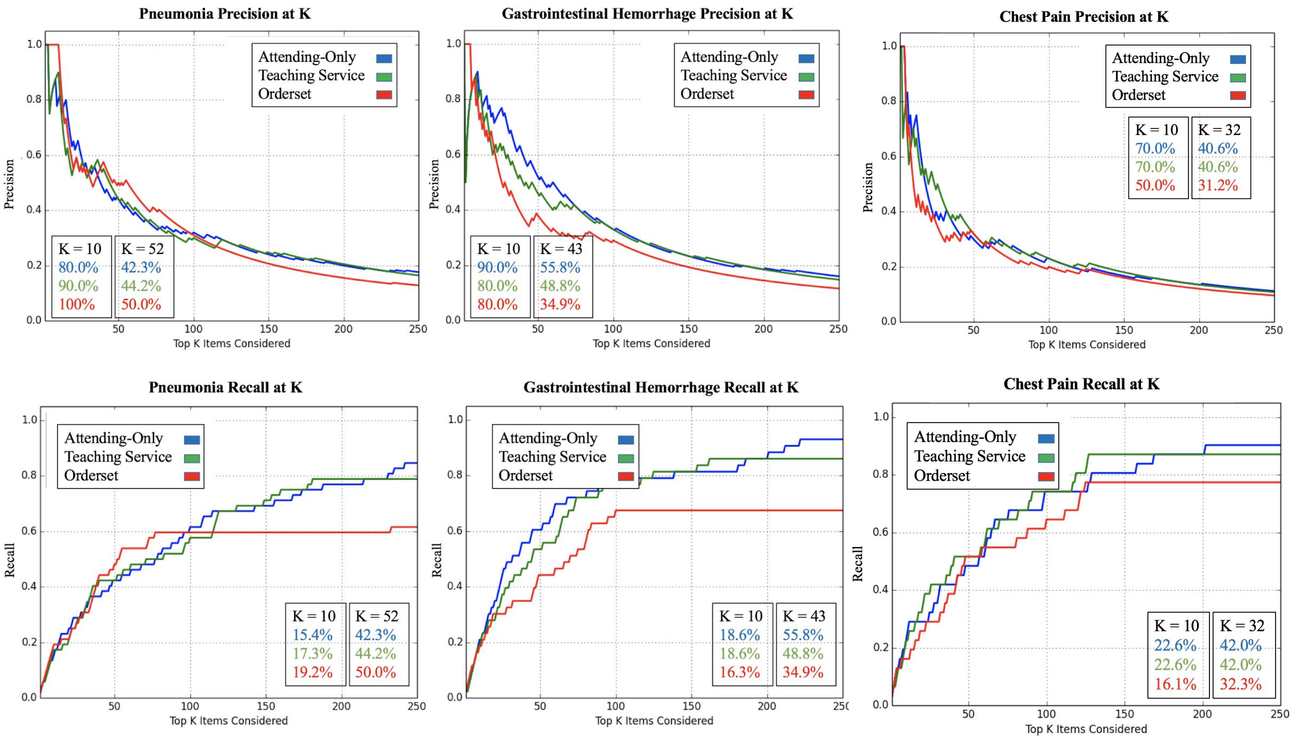


**Figure 3.** Precision (top) and recall (bottom) curves for 3 common diagnoses: pneumonia (ICD9: 486), gastrointestinal hemorrhage (ICD9: 578.9), and chest pain (ICD9: 786.5). Prediction accuracy (precision or recall) for predicting guideline reference orders is shown as a function of the top K recommendations considered (up to 250) using PPV as the scoring metric. Data labels are added for K = 10 and nO = number of items in the respective hospital order set. nO = 52, 43, and 32 for pneumonia, gastrointestinal hemorrhage, and chest pain, respectively. As the manually-curated hospital order set has no inherent ranking, orders are randomly sampled with replacement from the order set as the curve progresses from left to right.



## Discussion

In Table 1, we see that the top clinical orders predicted by attending-only and teaching service association models corroborate on several key items (e.g. levofloxacin, blood culture, CT head, etc.). Although the specific rankings and association statistics vary, as a discrete ranked list the order predictions yield comparable c-statistics (AUC) against the guideline reference standard with 95% confidence intervals satisfying the definition of "bioequivalence" across the 6 admission diagnoses (Figure 2). Both association models outperform hospital order sets for 5 of the 6 diagnoses, with $P<10^{-100}$ for all differences between predicted and benchmark c-statistics. In the case of heart failure, the teaching service model performed marginally worse. This instability can be attributed to small data size (n=38 for teaching service patients admitted under heart failure in 2013). In comparison, the teaching service cohort has substantially more (n>100) patients admitted for each of the 5 other diagnoses. Notably, in this hospital's teaching setting, heart failure patients are largely separated to a specialized heart failure teaching service away from the general medicine teaching service we investigate in this study. Thus, the distribution of patients and order patterns fed into the recommender algorithm may only capture an unusually distinct subset of heart failure admissions. In comparing the similarity between order lists predicted by the two patient cohorts, we see strong average overlap as indicated by RBO values of ~0.7 in Table 2. In relation to a prior study,[20] the attending-only and teaching service order lists across all 6 diagnoses share higher RBO values (greater stability and overlap) than learned order lists for 43 common admission diagnoses compared over time (e.g. order lists generated from 2009 versus 2012 EHR data). Precision and recall curves at varying values of K in Figure 3 pay particular attention to the top items that a clinician user could realistically be expected to review in a CPOE system. In the case of gastrointestinal hemorrhage and chest pain, we see that attending-only and teaching service order lists achieve greater recall and precision particularly at small values of K. These results support the argument that aggregating clinical order patterns from teaching services characterized by physician-trainee teams and attending-only services characterized exclusively by board-certified physicians will converge towards comparable top results as both cohorts share a common end goal of patient care. This study is an important step in addressing the concern of training machine-learning models on the the aggregate wisdom of clinicians, when the clinical setting contains both trainees and more experienced physicians.

Although propensity score matching was conducted to minimize the influence of confounding covariates recorded in EHR data, reducing the SMD between all recorded covariates to <0.1 as shown in Figure 1, learned clinical order patterns can be influenced by undocumented biases, not explicitly accounted for in demographic data, past medical history, or initial vital signs. In particular, existing hospital order sets are provided as a resource for clinicians to utilize. Automated association models may simply recapitulate the pre-authored order set templates in lieu of truly capturing individual clinical practice patterns. However, only 15% of all clinical orders recorded in the STRIDE dataset were input as part of an orderset.[44] Furthermore, in a supplementary sensitivity analysis, we compared association models trained with and without items input as part of hospital order sets for altered mental status, chest pain, and pneumonia. The resultant ROC curves were largely indistinguishable, providing reassurance that the learned practice patterns were not overly sensitive to or undermined by physicians using pre-existing order set templates. Propensity score matching was conducted on the full attending-only and teaching service patient cohorts. However, when given a query diagnosis (e.g. pneumonia), the clinical recommender engine considers only co-occurrence counts and subsequent association statistics between the specific diagnosis and all other candidate clinical items. Thus, even though the cohorts are balanced with respect to all patients, the subset of patients with a specific diagnosis may not be. Propensity score matching between the two cohorts on a per-diagnosis basis would further reduce potential bias.

In our previous work, we already demonstrated that these methods can accurately predict real-world clinical practice patterns.[19] However, a fair concern is whether real-world practices are synonymous with preferred ones. With the variability and uncertainty of medical practice, it can be perilous to define a gold standard to assess medical decision making. An important contribution of this study is that we evaluate our association models against an external reference standard of clinical practice guidelines to more closely align with patient outcomes and clinical trial evidence. The practice pattern learning methods easily extend to more varied clinical scenarios, but would not be appropriate for this study setting as we specifically focused on admission diagnoses with existing clinical practice guidelines and hospital order sets to enable our benchmark evaluations. Even clinical practice guidelines from trusted sources (e.g. National Guideline Clearinghouse) are published in the form of non-prescriptive clinical text that is deliberately open to interpretation. In this study, guideline reference standards were extracted by a single physician. To improve robustness of the curation process, reference standards can be generated by adjudicating lists generated by multiple physicians independently reviewing clinical practice literature. With the emergence of data-driven CDS systems rooted in machine-learning, it will become increasingly important to advance similar learning and evaluation methods as here for curating standard references for decision making quality. Indeed, clinical practice guidelines are imperfect. The fact that a clinical order was not mentioned in a guideline does not indicate it is an incorrect medical decision. However, defining "appropriate" orders for a given admission diagnoses is nearly impossible in the general case as there is no gold standard in medicine. As such, in this study and in future work, we seek to introduce a variety of different perspectives to evaluate machine-generated clinical order suggestions. Here, we ask whether experienced clinicians yield different practice patterns than trainees by comparing practice patterns learned from two distinct clinical settings. In future work, we focus on concrete patient outcomes by evaluating practice patterns learned from clinicians with substantially higher or lower observed versus expected patient mortality rates.

## Conclusion

When extracting patterns from the ongoing stream of practice data made available in EHR systems at academic hospitals, we are implicitly accepting the decision making of less experienced trainees alongside more experienced clinicians. Clinical order recommender systems trained on distinct patient cohorts, one seen exclusively by an attending-only service and the other seen by a teaching service characterized by physician-trainee teams, yield comparable top aggregate results that align with clinical practice guidelines as well as if not better than manually curated decision support content.

## References

1. Richardson WC, Berwick DM, Bisgard JC. Crossing the Quality Chasm: A New Health System for the 21st Century. National Academy Press. 2001. ISBN 0309072808.
2. Tricoci P, Allen JM, Kramer JM, et al. Scientific evidence underlying the ACC/AHA clinical practice guidelines. JAMA. 2009;831-841.
3. Durack DT. The weight of medical knowledge. N. Engl. J. Med. 1978;773-775.
4. 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;1409-1415.
5. Kawamoto K, Houlihan CA, Balas EA, et al. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ. 2005;765.
6. Ballard DW, Kim AS, Huang J, et al. Implementation of Computerized Physician Order Entry Is Associated With Increased Thrombolytic Administration for Emergency Department Patients With Acute Ischemic Stroke. Ann. Emerg. Med. 2015;1-10.
7. Ballesca MA, LaGuardia JC, Lee PC, et al. An electronic order set for acute myocardial infarction is associated with improved patient outcomes through better adherence to clinical practice guidelines. J. Hosp. Med. 2014;155-161.
8. Hripcsak G, Albers DJ, Perotte A. Exploiting time in electronic health record correlations. J. Am. Med. Informatics Assoc. 2011;109-115.
9. Sittig DF, Wright A, Osheroff JA, et al. Grand challenges in clinical decision support. J Biomed Inform. 2008;387-392.
10. Longhurst C, Harrington R, Shah NH. A “Green Button” For Using Aggregate Patient Data At The Point Of Care. Health Aff. 2014;1229-1235.
11. Doddi S, Marathe A, Ravi SS, et al. Discovery of association rules in medical data. Med. Inform. Internet Med. 2001;25–33.
12. Klann J, Schadow G, Downs SM. A method to compute treatment suggestions from local order entry data. AMIA Annu. Symp. Proc. 2010;387–91.
13. Klann J, Schadow G, McCoy JM. A recommendation algorithm for automating corollary order generation. AMIA Annu. Symp. Proc. 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. 2006;819–823.
15. Zhang Y, Padman R, Levin JE. Paving the COWpath: data-driven design of pediatric order sets. J. Am. Med. Inform. Assoc. 2014;304–311.
16. Klann JG, Szolovits P, Downs SM, et al. Decision support from local data: creating adaptive order menus from past clinician behavior. J. Biomed. Inform. 2014;84–93.
17. Wright AP, Wright AT, McCoy AB, et al. The use of sequential pattern mining to predict next prescribed medications. J. Biomed. Inform. 2014;73–80.
18. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. J. Biomed. Inform. 2010;891–901.
19. Chen JH, Podchiyska T, Altman RB. OrderRex: Clinical order decision support and outcome predictions by data-mining electronic medical records. J Am Med Informatics Assoc. 2016;339-348.
20. Linden G, Smith B, York J. Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Comput. 2003;76-80.
21. Surowiecki J. The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations. Doubleday, 2004. ISBN 0385503865.

22. Young JQ, Ranji SR, Wachter RM, et al. “July effect”: Impact of the academic year-end changeover on patient outcomes. Ann Intern Med. 2011;309-315.
23. Igwe E, Hernandez E, Rose S, et al. Resident participation in laparoscopic hysterectomy: impact of trainee involvement on operative times and surgical outcomes. Am J Obstet Gynecol. 2014;1-7.
24. Shaikh T, Wang L, Ruth K, et al, The impact of trainee involvement on outcomes in low-dose-rate brachytherapy for prostate cancer. Brachytherapy. 2016;156,162.
25. Lowe HJ, Ferris TA, Hernandez PM, et al. STRIDE--An integrated standards-based translational research informatics platform. In AMIA Annu Symp Proc. 2009;391-395.
26. Hernandez P, Podchiyska T, Weber S, et al, Automated mapping of pharmacy orders from two electronic health record systems to RxNorm within the STRIDE clinical data warehouse. AMIA Annu. Symp. Proc. 2009;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;32–37.
28. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat. 2011;150-161.
29. Chen JH, Altman RB. Automated physician order recommendations and outcome predictions by data-mining electronic medical records. AMIA Jt. Summits Transl. Sci. Proc. 2014;206-210.
30. Xiao H, Wang Y, Xu T, et al. Evaluation and treatment of altered mental status patients in the emergency department: Life in the fast lane. World J. Emerg. Med. 2012;270–277.
31. Skinner JS, Smeeth L, Kendall JM, et al. NICE guidance. Chest pain of recent onset: assessment and diagnosis of recent onset chest pain or discomfort of suspected cardiac origin. Heart. 2010;974–978.
32. Cooper A, Timmis A, Skinner J. Assessment of recent onset chest pain or discomfort of suspected cardiac origin: summary of NICE guidance. BMJ. 2010;340.
33. Accounting and Corporate Regulatory Authority (ACRA). Radiologic Management of Lower Gastrointestinal Bleeding. 2011;8-13.
34. Laine L, Jensen DM. Management of patients with ulcer bleeding. Am J Gastroenterol. 2012;345–360.
35. National Health Service (NHS). Acute upper gastrointestinal bleeding: management. 2012.
36. National Institute for Health and Care Excellence (NICE). Acute heart failure: diagnosis and management management. 2014.
37. Yancy CW, Jessup M, Bozkurt B, et al. 2013 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;240–327.
38. Mandell L, Wunderink RG, Anzueto A, 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;27-72.
39. Lim WS, Baudouin SV, George RC, et al. BTS guidelines for the management of community acquired pneumonia in adults: update 2009. BMJ Thorax. 2009;64;55.
40. Moya A, Sutton R, Ammirati F, et al. Guidelines for the diagnosis and management of syncope. Eur. Heart J. 2009;2631-2671.
41. Kendall MG. A New Measure of Rank Correlation. Biometrika. 1938;30;81–93.
42. Webber W, Moffat A, Zobel J. A similarity measure for indefinite rankings. ACM Trans. Inf. Syst. 2010;1-38.
43. Center for Drug Evaluation and Research (CDER). General Considerations. 2003;23.
44. Chen JH, Goldstein MK, Asch SM, et al. Dynamically evolving clinical practices and implications for predicting medical decisions. Pac Symp Biocomput. 2016.