DYNAMICALLY EVOLVING CLINICAL PRACTICES AND IMPLICATIONS FOR PREDICTING MEDICAL DECISIONS

JONATHAN H CHEN
Center for Innovation to Implementation (Ci2i), Veteran Affairs Palo Alto Health Care System
Palo Alto, CA 94304 USA
Center for Primary Care and Outcomes Research (PCOR), Stanford University
Stanford, CA 94305 USA
Email: jonc101@stanford.edu

MARY K GOLDSTEIN
Geriatrics Research Education and Clinical Center, Veteran Affairs Palo Alto Health Care System
Palo Alto, CA 94304 USA
Center for Primary Care and Outcomes Research (PCOR), Stanford University
Stanford, CA 94305 USA
Email: mary.goldstein@va.gov

STEVEN M ASCH
Center for Innovation to Implementation (Ci2i), Veteran Affairs Palo Alto Health Care System
Palo Alto, CA 94304 USA
Division of General Medical Disciplines, Department of Internal Medicine, Stanford University
Stanford, CA 94305 USA
Email: sasch@stanford.edu

RUSS B ALTMAN
Departments of Bioengineering and Genetics, Stanford University, Stanford, CA 94305 USA
Department of Medicine, Stanford University, Stanford, CA 94305 USA
Email: russ.altman@stanford.edu

Automatically data-mining clinical practice patterns from electronic health records (EHR) can enable prediction of future practices as a form of clinical decision support (CDS). Our objective is to determine the stability of learned clinical practice patterns over time and what implication this has when using varying longitudinal historical data sources towards predicting *future* decisions. We trained an association rule engine for clinical orders (e.g., labs, imaging, medications) using structured inpatient data from a tertiary academic hospital. Comparing top order associations per admission diagnosis from training data in 2009 vs. 2012, we find practice variability from unstable diagnoses with rank biased overlap (RBO)<0.35 (e.g., pneumonia) to stable admissions for planned procedures (e.g., chemotherapy, surgery) with comparatively high RBO>0.6. Predicting admission orders for future (2013) patients with associations trained on recent (2012) vs. older (2009) data improved accuracy evaluated by area under the receiver operating characteristic curve (ROC-AUC) 0.89 to 0.92, precision at ten (positive predictive value of the top ten predictions against actual orders) 30% to 37%, and weighted recall (sensitivity) at ten 2.4% to 13%, ($P<10^{-10}$). Training with more longitudinal data (2009-2012) was no better than only using recent (2012) data. Secular trends in practice patterns likely explain why smaller but more recent training data is more accurate at predicting future practices.

1. Introduction

Variability and uncertainty in medical practice compromise quality of care and cost efficiency, with overall compliance with evidence-based guidelines ranging from 20-80%.[1] Clinical decision support (CDS) tools, like order sets and alerts, reinforce best-practices by distributing information on relevant clinical orders (e.g., labs, imaging, medications),[2–5] but production is limited in scale by knowledge-based manual authoring of one intervention at a time by human experts.[6] If medical knowledge were fixed, manual approaches might eventually converge towards a comprehensive set of effective clinical decision support content from the top-down. The reality is instead a perpetually evolving body of knowledge that responds to new evidence, technology, and epidemiology that requires ongoing content maintenance to adapt to changing clinical practices.[7]

The meaningful use era of electronic health records (EHR)[8] creates an opportunity for data-driven clinical decision support (CDS) to reduce detrimental practice variability through the collective expertise of many practitioners in a learning health system.[9–13] Specifically, one of the "grand challenges" in CDS[14] is automated production of CDS from the bottom-up by data-mining clinical data sources. Such algorithmic approaches to clinical information retrieval could greatly expand the scope of medical practice addressed with effective decision support, and automatically adapt to an ongoing stream of evolving clinical practice data. This would fulfill the vision of a learning health system to continuously learn from real-world practices and translate them into usable information for implementation back at the point-of-care. The Big Data[13,15] potential of EHRs makes this vision possible, but the dynamic nature of clinical practices over time calls into question the presumption that learning from historical clinical data will inform future clinical practice. To fulfill the potential of real-time clinical prediction, we need to better understand how far back in time to mine EHRs while retaining predictive value for future decision making.

2. Background

To understand clinical practice patterns and inform potential decision support, we focus on the clinical orders (e.g., labs, imaging, medications) that concretely manifest point-of-care decision making. Prior research into data-mining for clinical decision support content includes use of association rules, Bayesian networks, and unsupervised clustering of clinical orders and diagnoses.[16–23] This prior research has largely ignored the temporal relationships between clinical data elements when training predictive models, treating individual patients or encounters as an unordered collection of items. In our own prior work, inspired by analogous information retrieval problems in recommender systems, collaborative filtering, and market basket analysis, we automatically generated clinical decision support content in the form of a clinical order recommender system[24] analogous to Netflix or Amazon.com's "Customer's who bought A also bought B" system.[25] This prior work[26] first examined the importance of matching the temporal relationship between clinical data elements to the respective timing of evaluation outcomes. For example, orders co-occuring within a short time period, such as the antibiotics vancomycin and piperacillin-tazobactam being ordered within one hour of each other, inform a more useful association than orders separated by several days of time. The impact of the temporal relationship between training and validation data has not been explored in this prior research (including our own). Instead, any evaluation of these predictive models was conducted

by separating patients into random train-test subsets. This is not representative of realistic applied scenarios however, where we would have to learn from historical clinical data to inform recommendations and predictions towards future patient encounters that have never previously occurred.

In this work, we seek to determine how varying longitudinal historical training data usage can impact prediction of future clinical practices. Furthermore, we seek to quantify which inpatient admission diagnoses exhibit the most stability vs. variability of clinical practice patterns over time.

3. Materials and Methods

We extracted deidentified patient data from the (Epic) electronic medical record for all inpatient hospitalizations at Stanford University Hospital via the STRIDE clinical data warehouse.[27] The structured data covers patient encounters from their initial (emergency room) presentation until hospital discharge. With five years of data spanning 2008-2014, the dataset includes >74K patients with >11M instances of >27K distinct clinical items. The clinical item elements include >7,800 medication, >1,600 laboratory, >1,100 imaging, and >1,000 nursing orders. Non-order items include >1,000 lab results, >7,800 problem list entries, >5,300 admission diagnosis ICD9 codes, and patient demographics. Medication data was normalized with RxNorm mappings[28] down to active ingredients and routes of administration. Numerical lab results were binned into categories based on "abnormal" flags established by the clinical laboratory. To compress the sparsity of diagnosis items, we duplicated ICD9 codes up to the three digit hierarchy, such that an item for code 786.05 would have additional items replicated for code 786.0 and 786. The above pre-processing models each patient as a timeline of clinical item 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,[29] most clinical items may be ignored with minimal information loss. In this case, ignoring rare clinical items with <256 instances reduces the effective item count from >27K to ~3K (11%), while still capturing 10.8M (95%) of the 11.4M item instances. After excluding common process orders (e.g., vital signs, notify MD, regular diet, transport patient, as well as most nursing and all PRN medications), 1,270 clinical orders of interest remained.

Using our previously described method,[24] we algorithmically mined association rules for clinical item pairs from past clinician behavior. Based on Amazon's product recommender,[25] we collected patient counts for all clinical item instance pairs co-occurring within 24 hours of each other to build time-stratified item association matrixes.[26] Each matrix defines a 2x2 contingency table for each pair of clinical items, from which various association statistics are derived (e.g., odds ratio (OR), positive predictive value (PPV), baseline prevalence, and Fisher's P-value).[30] To assess the varying impact of historical training data time, separate item association matrix models were built from training data from 2009, data from 2012, and data from 2009 through 2012.

We identify clinical order associations that reflect practice patterns by using query items (e.g., admission diagnosis or first several clinical orders and lab results) to score-rank all candidate clinical order items by an association statistic relative to the query items. Score-ranking by PPV (positive predictive value) prioritizes orders that are *likely* to occur after the

query items, while score-ranking by Fisher's P-value for items with odds ratio > 1 prioritizes orders that are *disproportionately associated* with the query items.[26]

To find clinical orders associated with different admission diagnoses, we generated a score-ranked list of the 1,270 candidate clinical orders for each of the most common admission diagnoses (those with at least 36 instances per year), sorted by Fisher's P-value. To assess for stability in clinical order patterns, we generated two such clinical order lists for each admission diagnosis, one from the matrix built on 2009 data and the other from the 2012 data. Traditional measures of list agreement like Kendall's τ[31] are not ideal here, as they often require identically sized, finite lists, and weigh all ranks equivalently. To compare ranked clinical order lists, we instead calculate their agreement by Rank Biased Overlap (RBO).[32] When comparing two ranked lists, we define I_k as the intersection of the top k items in each list, and X_k as the size of the overlap at rank depth = $|I_k|$. The ratio of X_k to the maximum possible value (k) is the fractional overlap agreement $A_k = (X_k/k)$. RBO is a weighted summation of these agreements where the weight $w_k = (1-p)*p^{k-1}$, based on the "persistence" parameter p that reflects the probability that an observer reviewing the top k items will continue to observe the ($k+1$)-th items. The fixed $(1-p)$ factor normalizes the sum of weights to 1. For our calculations, we used a default implementation p parameter of 0.98.[33]

$$RBO = \sum_{k=1}^{\infty} w_k \cdot A_k$$

The geometric weighting scheme of RBO serves to emphasize items at the top of the list and to ensure numerical convergence regardless of list length. RBO values range from 0.0 (disjoint lists) to 1.0 (identical lists).

To assess the utility of historical clinical item associations towards predicting future practices, we performed a variation of our prior experiments to predict hospital admission orders.[24] Specifically, using association matrices trained on data from 2009, 2012, or 2009 through 2012, we used the first four hours of clinical items from every future patient admitted to the hospital in 2013 to query for a ranked list of associated clinical orders. We compared these generated order lists against the actual next 24 hours of subsequent clinical orders (that did not already occur within the query time) by area under the receiver operating characteristic (ROC-AUC), precision (positive predictive value) at 10 items, and inverse frequency weighted recall[26] (sensitivity) at 10 items. Statistical tests (*t*-tests, Pearson's correlation) were calculated with the SciPy Python package.[34]

4. Results

Table 1 reports patient demographics and the flux of new and departing ordering providers in the clinical data over the years studied. Table 2a,b,c illustrate examples of the top associated clinical orders for different admission diagnoses based on 2009 vs. 2012 data, with corresponding calculations of ranked item overlap that define the Rank Biased Overlap (RBO) score for each pair of lists. Figure 1 depicts the Rank Biased Overlap (RBO) between 2009 vs. 2012 for each of the most common admission diagnoses. Figure 2 depicts the correlation between diagnosis stability (RBO) and accuracy towards predicting future order patterns (weighted recall). Table 3 reports the overall average accuracy metrics for predicting future (2013) clinical order patterns based on association matrices trained on different subsets of historical data (2009, 2012, or 2009 through 2012).

Table 1 – Patient demographics and provider flux over the evaluation period. New providers reflect those authorizing clinical orders during a given year, but not in the prior year. Similarly, departing providers reflect those from a given year, that are not found in the subsequent year.

Metric	2009	2010	2011	2012	2013
Patients	13,493	18,459	19,070	19,327	19,523
Age (mean)	58.4	58.1	58.6	58.5	58.6
Age (std dev)	18.5	18.8	18.7	18.7	18.6
Female	52%	52%	52%	51%	51%
White	60%	63%	62%	60%	58%
Hispanic/Latino	12%	13%	13%	13%	14%
Asian	11%	11%	11%	12%	12%
Black	5%	5%	5%	5%	5%
Providers	1,709	1,892	1,917	1,798	1,821
New Providers	…	41%	33%	29%	34%
Departing Providers	34%	32%	33%	33%	…

Table 2a – Top associated clinical orders for admission diagnosis of "Encounter for Chemotherapy" (ICD9: V58.11) based on 2009 and 2012 data, score-ranked by Fisher's P-value. At each rank *k*, the intersection of the top *k* items from each list (I_k) defines the "Overlap at Rank Depth" (X_k). The ratio of overlap to rank yields a "Fractional Overlap" agreement (A_k). For the full list of 1,270 candidate clinical orders, averaging the Fractional Overlap column with a geometric weighting scheme emphasizes the importance of the top items and ensures numerical convergence. The Rank Biased Overlap (RBO) score uses a weight for each A_k term, $w_k = (1\text{-}p)*p^{k-1}$, where *p* represents a "persistence" parameter reflecting the probability that the observer of *k* items is willing to continue to inspect the *k+1* items. RBO = 0.67 for this diagnosis, indicating relatively stable rankings compared to other diagnoses. This reflects standardized practices that have not significantly changed, including chemotherapeutic agents (cyclophosphamide, rituximab) and anticipatory co-medications for side effects (filgrastim for neutropenia; ondansetron, dexamethasone, aprepitant, and diphenhydramine for nausea).

2009 Top Items	Overlap at Rank Depth	Rank	Fractional Overlap	2012 Top Items
Cyclophosphamide (IV)	0	1	0.00	Ondansetron + Dexamethasone (IV)
Ondansetron + Dexamethasone (IV)	1	2	0.50	Aprepitant (Oral)
BMT Panel 1	1	3	0.33	Filgrastim (Subcutaneous)
Ondansetron (Oral)	2	4	0.50	Cyclophosphamide (IV)
BMT Panel 2	3	5	0.60	Ondansetron (Oral)
Rituximab (IV)	3	6	0.50	Dexamethasone (Oral)
Dexamethasone (Oral)	4	7	0.57	Diphenhydramine (Intravenous)
Aprepitant (Oral)	6	8	0.75	Rituximab (IV)
Filgrastim (Subcutaneous)	7	9	0.78	D5NS KCl NaAcetate Furosemide (IV)
Diphenhydramine (Intravenous)	8	10	0.80	D5NS KCl NaAcetate (IV)
…	…	…	…	…

Table 2b – Top associated clinical orders for admission diagnosis of "Pneumonia" (ICD9: 486) based on 2009 and 2012 data, score-ranked by Fisher's P-value. Rank Biased Overlap (RBO) = 0.35 between the two lists, indicating a substantial shift in the item rankings between the two lists. A dynamic change in practice patterns is evident in response to external, epidemiologic factors as 2009 saw much more testing (Respiratory DFA Panel, Influenza A PCR) and empiric treatment (Respiratory Isolation, Oseltamavir) for the H1N1 swine flu pandemic.[35,36] The viral pandemic dissipated by 2012, with the most prominent orders shifting towards empiric treatment for community acquired pneumonia[37] (azithromycin, ceftriaxone, levofloxacin) and antibiotic resistant organisms causing health care associated pneumonia[38] (vancomycin, piperacillin-tazobactam).

2009 Top Items	Overlap at Rank Depth	Rank	Fractional Overlap	2012 Top Items
Levofloxacin (IV)	0	1	0.00	Azithromycin (IV)
Blood Culture (2x Aerobic)	1	2	0.50	Levofloxacin (IV)
Blood Culture ((An)Aerobic)	1	3	0.33	Vancomycin (IV)
Respiratory DFA Panel	1	4	0.25	Piperacillin-Tazobactam (IV)
Respiratory Isolation	1	5	0.20	Ceftriaxone (IV)
Oseltamivir (Oral)	1	6	0.17	Azithromycin (Oral)
Vancomycin (IV)	2	7	0.29	Albuterol-Ipratropium (Inhalation)
Respiratory Culture	2	8	0.25	Sodium Chloride (Inhalation)
Albuterol-Ipratropium (Inhalation)	4	9	0.44	Blood Culture (2x Aerobic)
CBC w/ Diff	5	10	0.50	Blood Culture ((An)Aerobic)
Influenza A PCR	5	11	0.45	Ipratropium (Inhalation)
…	…	…	…	…

Table 2c - Top associated clinical orders for admission diagnosis of "Joint Pain" (ICD9: 719.4) based on 2009 and 2012 data, score-ranked by Fisher's P-value. Rank Biased Overlap (RBO) = 0.29 between the two lists, indicating a substantial shift in the item rankings between the two lists. Prominent orders in 2009 reflect diagnostic workup of arthritis (including fluid cell count and culture) while 2012 reveals more prominent symptomatic treatment with intravenous opioids (hydromorphone) that concomitantly require laxatives (sennosides, polyethylene glycol, magnesium citrate) to manage the predictable constipating side effects of opioids. The 2012 prominence of "Consult Orthopedics" suggests a shift in primary treatment teams from surgical to medical services since 2009.

2009 Top Items	Overlap at Rank Depth	Rank	Fractional Overlap	2012 Top Items
Overhead Bed Frame & Trapeze	0	1	0.00	Sennosides (Oral)
XR Pelvis 1V	0	2	0.00	Polyethylene Glycol (Oral)
Prothrombin TIME (PT/INR)	1	3	0.33	XR Pelvis 1V
CBC w/ Diff	1	4	0.25	Consult Orthopedics
Metabolic Panel, Basic	2	5	0.40	Overhead Bed Frame & Trapeze
XR Femur RT	2	6	0.33	Magnesium Citrate (Oral)
XR Shoulder 1V RT	2	7	0.29	Enoxaparin (Subcutaneous)
Cell Count, Synovial Fluid	2	8	0.25	XR Hip 2V LT
Fluid Culture and Gram Stain	2	9	0.22	Hydromorphone (Intravenous)
Bupivacaine (Nerve Block)	2	10	0.20	XR Femur LT
…	…	…	…	…

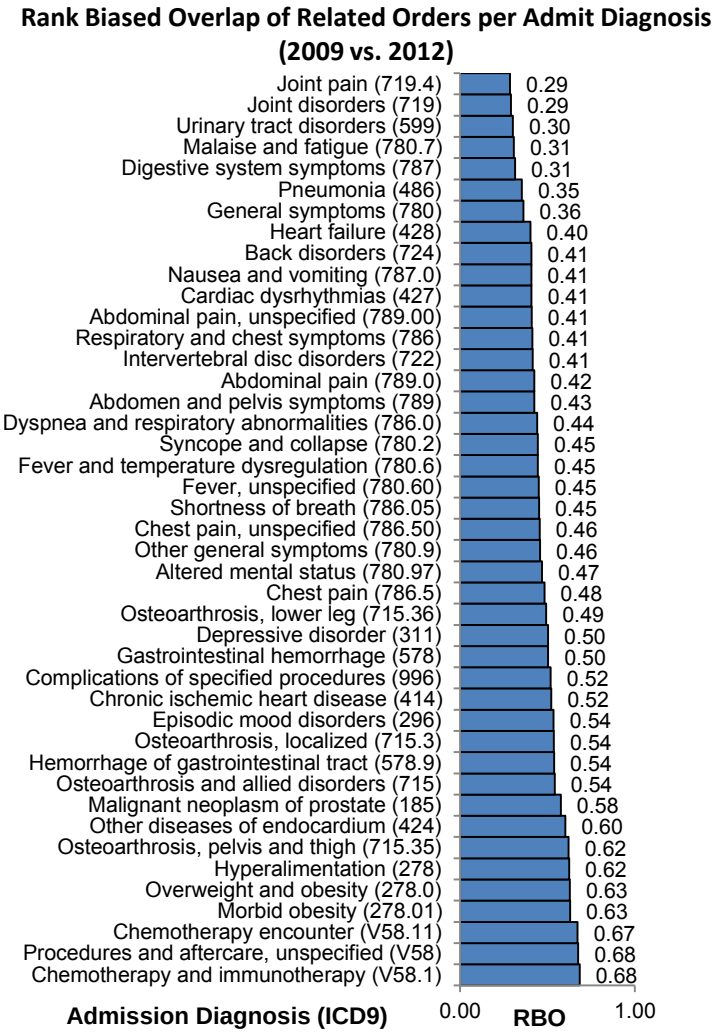

Figure 1 - Rank Biased Overlap (RBO) assessment of similarity of the lists of orders associated with the most common admission diagnoses in 2009 vs. 2012. Qualitative patterns reveal more stable ordering patterns (higher RBO) for elective hospital admissions with specific treatment plans and protocols like chemotherapy, obesity (sleep apnea and bariatric surgery), and osteoarthrosis (orthopedic surgery). Greater variability in ordering patterns (lower RBO) is seen for admission diagnoses with dynamically evolving practice patterns and less specific syndromes that may result in variable management, such as joint pain, malaise, and digestive symptoms.

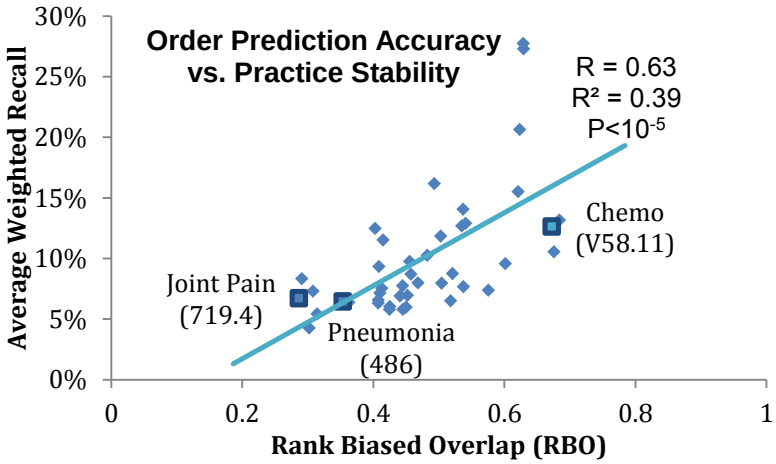

Figure 2 - Average weighted recall per admission diagnosis when predicting 2013 admission orders based on 2009-2012 training data by rank biased overlap. Using the association matrix trained on 2009-2012 data, the first clinical items from every admission in 2013 was used to query for the top ten associated clinical orders score-ranked by Fisher's P-value. Associated orders were compared against actual subsequent orders to yield a weighted recall score.[26] Each point represents the average weighted recall for one admission diagnosis vs. the respective rank biased overlap (RBO) score of order stability for 2009 vs. 2012. A linear trendline with Pearson correlation coefficient and two-tailed P-value illustrate a positive association between practice stability (higher RBO) and accuracy (weighted recall) towards predicting future order patterns.

Table 3 – Accuracy measures for predicting 2013 admission orders when using training data from different subsets of prior years. For ~15K patients with ~26K hospital admissions in 2013, data from the first four hours for each admission was used to query an association matrix trained on prior year(s) data for a list of clinical orders. The list of generated orders is score-ranked by PPV (positive predictive value ~ post-test probability) to identify orders *likely* to occur or by P-value to prioritize orders *disproportionately associated* with the query items. Generated order lists were compared against the subsequent 24 hours of clinical orders that actually occurred in each 2013 admission. Full list ranking is evaluated by the area under the receiver operating characteristic curve (ROC-AUC), while precision at ten evaluates only the top ten items. Inverse frequency weighted recall identifies methods most effective at retrieving less common, but specifically relevant orders.[26] Compared against the 2013 results, all bolded average results differed with $P<10^{-5}$ by two-tailed paired *t*-tests.

Training Data Year(s)	Training Patients	Average ROC-AUC PPV Associations	Average Precision at Ten PPV Associations	Average Weighted Recall at Ten P-value Associations
2009	10,727	0.888	**29.8%**	**2.4%**
2012	12,503	0.922	36.8%	**13.4%**
2011-2012	21,901	0.921	36.1%	**12.9%**
2009-2012	34,812	0.919	**35.4%**	**12.2%**
2013	11,278	0.924	38.0%	16.5%

5. Discussion

These results support the general supposition that clinical practices dynamically change over time (Figure 1). Elective admissions for planned procedures like chemotherapy and surgeries appear to exhibit relatively less variability over time with higher RBOs. This could of course be disrupted if future practices shifted in response to newly discovered different chemotherapy or surgical regimens, though the identified associations could still be reasonably used to suggest co-medications that are not enforced through a strict protocol. Diagnoses subject to epidemiologic shifts (i.e., pneumonia) and medical admissions for non-specific symptoms (e.g., joint pain, malaise) may trigger variable approaches to workup, represented by their lower RBOs. This method provides a quantitative assessment of clinical practice areas with the most dynamic changes, with respective implications on the reproducibility and reliability of predicting future clinical practice patterns based on historical data. It also has implications for ongoing debates on the appropriate interval for continuing medical education and maintenance of certification for individual clinicians.[39,40] For example, it could be used to identify areas where frequent education is required to adapt to rapidly shifting standards of practice vs. areas with years of stable practices that diminish the value of repetitious education maintenance.

Table 3 Table 3 reports the accuracy of models trained on different subsets towards predicting future practices by multiple measures. The area under the ROC curve (ROC-AUC) assesses discrimination accuracy for the full ranked list of candidate orders. Precision at ten items pays particular attention to the top items that a human user could realistically be expected to review. Weighted recall highlights retrieval of more "interesting" and specifically relevant suggestions over common, but potentially mundane, suggestions.[26] As might be expected, clinical order recommenders trained on more recent (2012) data are more accurate at predicting future (2013) practices than older (2009) data by all measures. The more compelling question answered is whether training on a larger longitudinal dataset (2009-2012) yields better results than just using the most recent data (2012). In this case, the extended data set is no better to slightly worse than just using the most recent data. While larger datasets are generally expected to improve the power of statistical learning methods, the correlation with RBO in Figure 2 suggests the changing clinical practice patterns over time makes older data less relevant when predicting future events.

This study focuses on the relevance of learned clinical order patterns towards predicting future events, but provides no assurance that common or strongly associated behaviors actually reflect "good" decisions. Short of randomized trials, we are evaluating our order associations against the external standards-of-care established in clinical practice guidelines.[43] With the results of this study however, it is not surprising that practice guidelines themselves must undergo regular revision, resulting in an ambiguous and moving target of clinical decision making quality that defies the existence of a fixed gold standard for clinical decision support.

A potential limitation in our evaluation of clinical practice pattern stability is the presumption that changing patterns reflect changes in clinical decision making at the management and treatment level. The nature of the EHR data source likely results in changing order patterns due to non-clinical data changes, such as shifts in diagnosis coding practices from pneumonia to sepsis.[41] Administrative infrastructure changes are expected to occur despite having little semantic difference for clinical decision making, such as the hospital orders for Respiratory Virus DFA (direct fluorescent antibody) panels being replaced with Respiratory Virus PCR (polymerase chain reaction) panels. Related work we are undertaking on probabilistic topic models of clinical data could provide opportunities to detect and resolve such "semantic" differences by noting that both such respiratory virus tests are related to "respiratory infection" scenarios, even though the two are never found together for a single patient. There may also be a substantial shift in patient characteristics insufficiently captured by admission diagnosis stratification, such as patient admissions for "joint pain" that might represent anything from elective orthopedic surgery admissions, workup for suspected septic arthritis, to pain management for a rheumatoid arthritis flare. Using more robust cohort identification methods than admission ICD9 codes, such as through natural language processing of clinical notes or SNOMED-CT codes could help normalize such factors. Individual patients could be hospitalized multiple times within each evaluation period, which could bias the association statistics without clustering statistics to mitigate internally correlated data. With all data deriving from a single medical center, significant cultural shifts in practice patterns could also be unduly influenced by the large flux of providers noted in Table 1 or even a small number of prominent clinicians.

Even if learned clinical practice patterns change for "non-clinical" reasons above, the overarching caution of depending on historical data to predict future clinical events remains

relevant. The evolving clinical patterns reinforce the challenge of manually producing clinical decision support and knowledge guides for order entry, as they must be followed by ongoing manual effort to maintain them against new clinical evidence and standards that may substantially shift within just a few years. Automated algorithms to learn clinical decision support are thus even more important to not only cover the breadth of medical knowledge efficiently, but to automatically adapt to continuous streams of new information. While historical data will not predict the advent of new therapeutics or diseases, incorporating a continuous stream of data could allow automated methods to rapidly detect and adapt to shifting practice changes and alert authors to dynamic areas in need of additional decision support, just as Google Flu Trends can detect local flu activity more rapidly than conventional methods.[42] The results above inform such an approach, indicating that using the most recent data may be more important than simply accumulating a massive repository of historical data whose interpretation does not even remain internally consistent. Future opportunities could explore weighted or online learning algorithms that emphasize the relevance of recent data without completely ignoring the older data that may still capture useful information.

6. Conclusions

Clinical practice patterns for hospital admission diagnoses (automatically) learned from historical EHR data can vary substantially across years, particularly for non-specific symptom-based diagnoses and those influenced by external epidemiology (e.g., pneumonia). Elective admissions for planned procedures (e.g., chemotherapy, surgery) demonstrate more stable practice patterns over time. If the goal is predicting relevant future practices, using more recent training data is more accurate than using older data, likely due to secular trends in changing practice. Consequently, using a larger longitudinal data set from many years may be no better, and possibly worse, than using a smaller but more recent data set. Decision support and predictive analytic models should take these patterns into account.

7. Acknowledgments

Project supported by the Stanford Translational Research and Applied Medicine (TRAM) program in the Department of Medicine (DOM) 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 (NCATS) 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 Lee Ann Yasukawa and Susan Weber 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.

Views expressed are those of the authors and do not necessarily represent the official views of the NIH, Department of Veteran Affairs (VA) or other affiliated institutions.

References

1. Richardson, W. C. *et al. Crossing the Quality Chasm: A New Health System for the 21st Century*. *Natl. Acad. Press* (Institute of Medicine, Committee on Quality of Health Care in America Committee on Quality of Health Care in America, 2001). doi:10.1136/bmj.323.7322.1192
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. *Arch. Intern. Med.* **163,** 1409–1416 (2003).
3. Overhage, J. & Tierney, W. A randomized trial of 'corollary orders' to prevent errors of omission. *J. Am. Med. Informatics Assoc.* **4,** 364–75 (1997).
4. Tierney, W. M. *et al.* Computerizing Guidelines to Improve Care and Patient Outcomes: The Example of Heart Failure. *J. Am. Med. Informatics Assoc.* **2,** 316–322 (1995).
5. Chen, J. H. *et al.* Why providers transfuse blood products outside recommended guidelines in spite of integrated electronic best practice alerts. *J. Hosp. Med.* (2014). doi:10.1002/jhm.2236
6. Bates, D. W. *et al.* Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *J. Am. Med. Inform. Assoc.* **10,** 523–30 (2003).
7. Goldstein, M. K. *et al.* Implementing clinical practice guidelines while taking account of changing evidence: ATHENA DSS, an easily modifiable decision-support system for managing hypertension in primary care. *Proc. AMIA Symp.* 300–4 (2000). at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2243943&tool=pmcentrez&rendertype=abstract>
8. 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.* **77,** 54163–292 (2012).
9. Longhurst, C. a., Harrington, R. a. & Shah, N. H. A 'Green Button' For Using Aggregate Patient Data At The Point Of Care. *Health Aff.* **33,** 1229–1235 (2014).
10. Frankovich, J., Longhurst, C. A. & Sutherland, S. M. Evidence-based medicine in the EMR era. *N. Engl. J. Med.* **365,** 1758–9 (2011).
11. Smith, M., Saunders, R., Stuckhardt, L. & McGinnis, J. M. *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). doi:10.5860/CHOICE.51-3277
12. Krumholz, H. M. Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. *Health Aff.* **33,** 1163–1170 (2014).
13. De Lissovoy, G. Big data meets the electronic medical record: a commentary on 'identifying patients at increased risk for unplanned readmission'. *Med. Care* **51,** 759–60 (2013).
14. Sittig, D. F. *et al.* Grand challenges in clinical decision support. *J. Biomed. Inform.* **41,** 387–92 (2008).
15. Bates, D. W., Saria, S., Ohno-Machado, L., Shah, a. & Escobar, G. Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. *Health Aff.* **33,** 1123–1131 (2014).
16. Doddi, S., Marathe, a, Ravi, S. S. & Torney, D. C. Discovery of association rules in medical data. *Med. Inform. Internet Med.* **26,** 25–33 (2001).
17. Klann, J., Schadow, G. & Downs, S. M. A method to compute treatment suggestions from local order entry data. *AMIA Annu. Symp. Proc.* **2010,** 387–91 (2010).
18. Klann, J., Schadow, G. & McCoy, J. M. A recommendation algorithm for automating corollary order generation. *AMIA Annu. Symp. Proc.* **2009,** 333–7 (2009).
19. 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 Annu. Symp. Proc.* **2006,** 819–823 (2006).
20. Zhang, Y., Padman, R. & Levin, J. E. Paving the COWpath: data-driven design of pediatric order sets. *J. Am. Med. Inform. Assoc.* **21,** e304–e311 (2014).
21. Klann, J. G., Szolovits, P., Downs, S. M. & Schadow, G. Decision support from local data: creating adaptive order menus from past clinician behavior. *J. Biomed. Inform.* **48,** 84–93 (2014).
22. Wright, A. P., Wright, A. T., McCoy, A. B. & Sittig, D. F. The use of sequential pattern mining to predict next prescribed medications. *J. Biomed. Inform.* **53,** 73–80 (2014).
23. Wright, A., Chen, E. S. & Maloney, F. L. An automated technique for identifying associations between medications, laboratory results and problems. *J. Biomed. Inform.* **43,** 891–901 (2010).

24. Chen, J. H. & Altman, R. B. Mining for clinical expertise in (undocumented) order sets to power an order suggestion system. *AMIA Summits Transl. Sci. Proc.* **2013,** 34–8 (2013).
25. Linden, G., Smith, B. & York, J. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Comput.* **7,** 76–80 (2003).
26. Chen, J. H. & Altman, R. B. Automated Physician Order Recommendations and Outcome Predictions by Data-Mining Electronic Medical Records. in *AMIA Summits Transl. Sci. Proc.* 206–210 (2014).
27. Lowe, H. J., Ferris, T. a, Hernandez, P. M. & Weber, S. C. STRIDE--An integrated standards-based translational research informatics platform. *AMIA Annu. Symp. Proc.* **2009,** 391–5 (2009).
28. 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,** 244–8 (2009).
29. Wright, A. & Bates, D. W. Distribution of Problems, Medications and Lab Results in Electronic Health Records: The Pareto Principle at Work. *Appl. Clin. Inform.* **1,** 32–37 (2010).
30. Finlayson, S. G., Lependu, P. & Shah, N. H. Building the graph of medicine from millions of clinical narratives. *Sci. Data* **1,** 140032 (2014).
31. Kendall, M. G. A New Measure of Rank Correlation. *Biometrika* **30,** 81–93 (1938).
32. Webber, W., Moffat, A. & Zobel, J. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.* **28,** 1–38 (2010).
33. Agrawal, R. Comparing Ranked List. (2013). at <https://ragrawal.wordpress.com/2013/01/18/comparing-ranked-list/>
34. Jones, E., Oliphant, T., Peterson, P. & Al, E. SciPy: Open source scientific tools for Python. at <http://www.scipy.org>
35. Kerr, J. R. Swine influenza. *J. Clin. Pathol.* **62,** 577–578 (2009).
36. Who. WHO Guidelines for Pharmacological Management of Pandemic Influenza A(H1N1) 2009 and other Influenza Viruses. *WHO Guidel. Pharmacol. Manag. Pandemic Influ. A(H1N1) 2009 other Influ. Viruses* 1–32 (2010). at <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:WHO,2010#7>
37. Mandell, L. 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.* **44 Suppl 2,** S27–72 (2007).
38. Focaccia, R. & Gomes Da Conceicao, O. J. Guidelines for the management of adults with hospital-acquired, ventilator-associated, and healthcare-associated pneumonia. *Am. J. Respir. Crit. Care Med.* **171,** 388–416 (2005).
39. Teirstein, P. Boarded to Death — Why Maintenance of Certification Is Bad for Doctors and Patients. *N Engl J Med* **372,** 106–8 (2015).
40. Irons, M. B. & Nora, L. M. Maintenance of Certification 2.0 — Strong Start, Continued Evolution. *N. Engl. J. Med.* **372,** 104–106 (2015).
41. Rhee, C., Gohil, S. & Klompas, M. Regulatory mandates for sepsis care--reasons for caution. *N. Engl. J. Med.* **370,** 1673–1676 (2014).
42. Ginsberg, J. *et al.* Detecting influenza epidemics using search engine query data. *Nature* **457,** 1012–1014 (2009).
43. Chen, J. H. & Altman, R. B. Data-Mining Electronic Medical Records for Clinical Order Recommendations : Wisdom of the Crowd or Tyranny of the Mob ? *AMIA Summits Transl. Sci. Proc.* (2015).
44. Brin, S. & Page, L. The anatomy of a large-scale hypertextual Web search engine BT - Computer Networks and ISDN Systems. **30,** 107–117 (1998).
45. Excell, D. Bayesian inference - The future of online fraud protection. *Comput. Fraud Secur.* **2012,** 8–11 (2012).