Published in final edited form as:
Int J Med Inform. 2017 June ; 102: 71–79. doi:10.1016/j.ijmedinf.2017.03.006.

Decaying Relevance of Clinical Data Towards Future Decisions in Data-Driven Inpatient Clinical Order Sets

Jonathan H Chen, MD, PhD[1], **Muthuraman Alagappan, MD**[4], **Mary K Goldstein, MD, MS**[2,3], **Steven M Asch, MD, MPH**[1,6], and **Russ B Altman, MD, PhD**[1,5]

[1]Department of Medicine, Stanford University, Stanford, CA, USA

[2]Geriatrics Research Education and Clinical Center, Veteran Affairs Palo Alto Health Care System, Palo Alto, CA, USA

[3]Primary Care and Outcomes Research (PCOR), Stanford University, Stanford, CA, USA

[4]Internal Medicine Residency Program, Beth Israel Deaconess Medical Center, Boston, MA USA

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

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

Abstract

Objective—Determine how varying longitudinal historical training data can impact prediction of future clinical decisions. Estimate the “decay rate” of clinical data source relevance.

Materials and Methods—We trained a clinical order recommender system, analogous to Netflix or Amazon’s “Customers who bought A also bought B…” product recommenders, based on a tertiary academic hospital’s structured electronic health record data. We used this system to predict future (2013) admission orders based on different subsets of historical training data (2009 through 2012), relative to existing human-authored order sets.

Corresponding Author: Jonathan H. Chen MD, PhD, jonc101@stanford.edu, Telephone: 650-721-6669, Address: 1265 Welch Road, Stanford, CA 94305.

Role of the Funder/Sponsor: The sponsors had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript. Content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, VA or Stanford Healthcare.

Author Contributions: Dr. Chen had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: All authors
Acquisition of data: Chen
Analysis and interpretation of data: Chen, Alagappan
Drafting of the manuscript: Chen
Critical revision of the manuscript for important intellectual content: All authors
Statistical analyses: Chen, Alagappan
Administrative, technical, or material support: Altman, Asch
Study supervision: Altman, Goldstein, Asch

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Results—Predicting future (2013) inpatient orders is more accurate with models trained on just one month of recent (2012) data than with 12 months of older (2009) data (ROC AUC 0.91 vs. 0.88, precision 27% vs. 22%, recall 52% vs. 43%, all $P<10^{-10}$). Algorithmically learned models from even the older (2009) data was still more effective than existing human-authored order sets (ROC AUC 0.81, precision 16% recall 35%). Training with more longitudinal data (2009–2012) was no better than using only the most recent (2012) data, unless applying a decaying weighting scheme with a "half-life" of data relevance about 4 months.

Discussion—Clinical practice patterns (automatically) learned from electronic health record data can vary substantially across years. Gold standards for clinical decision support are elusive moving targets, reinforcing the need for automated methods that can adapt to evolving information.

Conclusions and Relevance—Prioritizing small amounts of *recent* data is more effective than using larger amounts of older data towards future clinical predictions.

Keywords

Electronic Health Records; Data Mining; Collaborative Filtering; Practice Variability; Prediction Models

1 INTRODUCTION

1.1 Background and Significance

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] Even after current reforms,[2] evidence-based medicine from randomized controlled trials cannot keep pace with the perpetually expanding breadth of clinical questions, with only ~11% of guideline recommendations backed by high quality evidence.[3] Clinicians are left to synthesize vast streams of information for each individual patient in the context of a medical knowledge base that is both incomplete and yet progressively expanding beyond the cognitive capacity of any individual.[4,5] The practice of medicine is thus routinely driven by individual expert opinion and anecdotal experience.

Clinical decision support (CDS) seeks to reinforce best-practices by distributing knowledge-based content through order sets, alerts, templates, and prognosis scoring systems.[6–10] Here we pay special attention to clinical orders (e.g., labs, imaging, medications) as the concrete manifestation of point-of-care decision making. Computerized provider order entry (CPOE)[11] typically occurs on an "a la carte" basis where clinicians search for and enter orders to trigger subsequent clinical actions (e.g., pharmacy dispensing and nurse administration of a medication, or phlebotomy collection and laboratory analysis of blood tests). Because clinician memory and intuition can be error-prone, health systems produce order set templates as a common mechanism to distribute standard practices and knowledge (in paper and electronic forms) as the current standard for executable clinical decision support. Clinicians can search by keyword for common scenarios (e.g., "pneumonia"), and hope they find a preconstructed order set that includes relevant orders (e.g., blood cultures, antibiotics, chest X-rays).[12–14]

While existing approaches to clinical decision support can already reinforce consistency with best-practices,[6,7,15–18] production of this content is limited in scale by the human-expert, knowledge-based authoring necessary for each intervention.[19] If medical knowledge were static, such 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 practice of medicine that responds to new evidence, technology, epidemiology, and culture that requires ongoing content maintenance to adapt to changing clinical practices.[20–22]

The meaningful use era of electronic health records (EHR)[23] creates an opportunity for data-driven clinical decision support (CDS) to reduce detrimental practice variability with the collective expertise of many practitioners in a learning health system.[24–28] Specifically, one of the “grand challenges” in clinical decision support is data-mining content from the bottom-up in clinical data sources.[29] Such algorithmic approaches to clinical information retrieval could greatly expand the scope of medicine addressed with effective decision support, and automatically adapt to an ongoing stream of evolving practice data. This would fulfill the vision of a health system that continuously learns from real-world practices and translates them into usable information for implementation back at the point-of-care. Prior research into data-mining for decision support content includes association rules, Bayesian networks, and unsupervised clustering of clinical orders and diagnoses.[30–37] In our own prior work, inspired by analogous information retrieval problems in collaborative filtering and market basket analysis, we produced a clinical order recommender system[38,39] analogous to Netflix or Amazon.com’s “Customer’s who bought A also bought B” system.[40]

Accumulating data in EHRs makes these concepts possible, but the dynamic nature of clinical practices over time challenges the presumption that learning from historical clinical data will inform current and future clinical practices. Prior work already demonstrates the importance of temporal patterns between clinical events towards outcome predictions.[39,41–43] Another important relationship is the separation between when data is generated relative to the time learned prediction models are applied and evaluated. Prior clinical prediction modules from mortality risk scores like APACHE and SAPS[44] to hospital readmissions models that risk adjust quality indicators[45] to modern systems based on electronic medical record data[10,46,47] all tend to evaluate their utility by assessing prediction accuracy on a (randomly) separated validation subset of the same data source. This is not representative of a realistic applied scenario where we must make recommendations and predictions towards *future* events that have not yet occurred.[48]

1.2 Objective

To determine how varying longitudinal historical training data usage can impact prediction of future clinical decisions. Determine which inpatient admission diagnoses exhibit the most stability vs. variability of clinical practice patterns over time. Estimate the “decay rate” of the relevance of clinical data sources for informing future predictions.

2 MATERIALS AND METHODS

2.1 Collaborative Filtering for Clinical Order Decision Making

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.[49] 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 >55M instances of >45K distinct clinical items. The clinical item elements include >10,000 medication, >1,600 laboratory, >1,200 imaging, and >1,000 nursing orders. Non-order items include >7,000 lab results, >7,800 problem list entries, >5,300 admission diagnosis ICD9 codes, and patient demographics. Medication data was normalized with RxNorm mappings[50] 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. We aggregated ICD9 codes up to the three digit hierarchy as in Table 1. This helps compress the sparsity of diagnosis categories, while retaining the original detailed codes if they are sufficiently prevalent to be useful. The above pre-processing models 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,[51] most item types may be ignored with minimal information loss. In this case, ignoring rare clinical items with <256 instances reduces the effective item count from >45K to ~4.6K (10%), while still capturing 54.5M (98%) of the 55.4M 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), 2,030 clinical orders remain.

Using our previously described method,[38,39,52] we algorithmically mined temporal association rules for clinical item pairs from past clinician behavior. Based on Amazon’s product recommender,[40] 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.[52] For pairs of items A and B, the co-occurrence counts accumulated can be represented as $N_{AB,t}$: Number of patients for whom item B follows A within time t, as illustrated in the pseudocode below. For each pair of clinical items, we use these counts to populate 2x2 contingency tables from which association statistics are derived (e.g., odds ratio (OR), positive predictive value (PPV), baseline prevalence, and P-value by chi-square test with Yates’ correction).[39,53]

```
Pseudocode: Item Co-occurrence Counting
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 ($t_B - t_A$) <= timeBin and (P,A,B) newly
```

```
encountered:
                                             Increment $N_{AB,timeBin}$
                   Record (P,A,B) as previously encountered
          Record (P,A,$t_A$) as previously analyzed
```

We identify clinical order associations that reflect practice patterns from the training data 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)[54] 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.[52]

2.2 Assessing Stability of Ordering Patterns

To find clinical orders associated with different admission diagnoses, we generated a score-ranked list of the 2,030 candidate clinical orders for each admission diagnosis, sorted by P-value. To assess for stability in these 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 2012 data. Traditional measures of list agreement like Kendall's **τ**[55] are not ideal to here, as they often require identically sized, finite lists, and weigh all list positions equivalently. To compare each pair of ranked clinical order lists, we instead calculate their agreement by Rank Biased Overlap (RBO),[56] with an implementation *p* parameter of 0.98.[57] RBO counts the average fraction of top items in common between the two ordered lists, geometrically weighted to emphasize the top of the list and to ensure numerical convergence regardless of list length. RBO values range from 0.0 (no correlation or random list order) to 1.0 (perfect agreement of list order).

Similarly, we assessed the top clinical orders associated with assignment of a Primary or Consulting treatment team (e.g., medicine vs. surgery vs. neurology) to patient care. Orders associated with specific teams were based on orders occurring within 24 hours of assignment, sorted by P-value.

2.3 Assessing Prediction of Clinical Orders

We assessed the ability of the learned item association rules to predict subsequent clinical orders as a variation of our prior experiments to predict hospital admission orders.[38] For a separate random selection of 4,820 (~25%) validation patients from 2013, we isolated each use of a pre-existing human-authored order set within the first 24 hours of each hospitalization. Using the human-authored order sets, up to date as of that moment of their use, provides a real-world reference point of prediction accuracy. For an order set consisting of *K* suggested items, we evaluated them against the "correct" set of orders that actually occurred for the patient within 24 hours in terms of precision (positive predictive value) and recall (sensitivity) at *K*.

We simulated production of an individually personalized, association rule-based "order set" at each moment in time when a real order set was used. To dynamically generate this content, the system evaluates the patient's available clinical data up to that point to score-rank a list of suggested orders by estimated PPV (positive predictive value ~ post-test probability). We evaluated the full score-ranked list of clinical order suggestions by area under the receiver operating characteristic curve (ROC AUC = c-statistic), and the top K suggestions in terms of precision and recall at K for predicting subsequent orders.

In previous work, we performed sensitivity analyses with respect to the follow-up verification time to look for “correct” orders and found a predictable pattern of increased precision (positive predictive value) but reduced recall (sensitivity) as the follow-up time was extended from 1 minute to 24 hours.[58] For the purposes of this study, evaluating a single common time point is sufficient to validate the relative retrieval rates of different training methods.

2.3 Assessing Historical Time Variation Effect on Future Prediction Accuracy

To assess the varying impact of historical training data time, we trained multiple item association models from different training time periods for a separate randomly selected subset of 41,127 training patients. Specifically, one matrix was trained on the 12 months of data from 2013 (typical random train-test split), one from the 12 months of data from 2009 (“old” data), and a series of models counting backwards from the end of 2012 (“recent” data) for different windows of training time from 1 to 48 months. For example, the 2012 based model trained on one month used data from 12/1/2012 to 12/31/2012, while the 48 month model used data from 1/1/2009 to 12/31/2012. Statistical *t*-tests were calculated with the SciPy Python package.[59]

To avoid a stark and somewhat arbitrary cut-off between “old” vs. “recent” windows of time, we tried several variations of using 48 months of historical training data with a smooth decaying weighting scheme to prioritize recent data.[60] To do so, we applied an exponentially decaying weight to data based on how old they are relative to the last time. This is elaborated further based on the definitions and pseudocode below:

- d: Delta / interval length of time during which parts of a model are trained.
- N: Number of occurrences of an item (pair), the basis for all association rule statistics.
- N_i: Number of occurrences of an item (pair) during the i-th interval.
- t_0: Initial time of model training.
- t_z: Final time of model training.
- z: Total number of intervals evaluated up to time t_z.
- w: Window time of interest
    - For the decaying model, data counts at time interval i are decayed by a weighting factor of $(1-\frac{1}{w})^{z-i}$. This has the effect of data at time t_z-w being weighted down by ~1/e, where e is the base of the natural logarithm.

  - For the non-decaying models, this would reflect a simple truncated "sliding window" where any data occurring before t_z-w is discarded (given a weight of zero) while all data occurring within (t_z-w, t_z) is given a full weight of one.

- $t_{1/2} = w \ln 2$: "Half-life" of data relevance in the decaying model. Corresponds to w representing the "mean lifetime" of data relevance.
- $\hat{N}=\sum_{i=1}^{z} N_i (1-\frac{1}{w})^{z-i}$: Decaying windows estimate for N.

Algorithmically, the summation notation for count estimates above can be calculated by sequentially accumulating counts and multiplying those numbers by ($1-\frac{1}{w}$) after each interval of evaluation time. Such a streaming approach (as outlined in the pseudocode below) has the additional benefit of being able to adapt to a continuous stream of input into perpetuity without having to reset the training process. This further reflects the intuitive advantage of using an exponential decay model as it allows for incorporation of an arbitrarily (even infinitely) long historical data record while still ensuring numerical convergence.

```
Pseudocode: Decaying Item Co-occurrence Counting
Initialize N count estimates = 0
Set $t_i$ = $t_0$
Set $t_{i+1}$ = $t_i$ + d
while $t_i$ < $t_z$:
         Run Item Co-Occurrence Counting from $t_i$ to $t_{i+1}$ to generate $N_i$ estimates
         Update all N = N * (1-1/w)
         Update all N = N + $N_i$
         Set $t_i$ = $t_{i+1}$
         Set $t_{i+1}$ = $t_i$ + d
```

3 RESULTS

Table 2 illustrates examples of the top clinical order associations for an example admission diagnosis of pneumonia, based on 2009 vs. 2012 data. Corresponding calculations of ranked item overlap that define the Rank Biased Overlap (RBO) score are included to reflect the relative stability vs. dynamic change in the ranked lists. The example illustrates standard workup and treatment for pneumonia (e.g., blood cultures, levofloxacin, venous blood gas + lactate) as prominent in both time periods. However, a dynamic change is evident in response to epidemiologic factors as 2009 saw much more testing (Respiratory Virus Direct Fluorescent Antibody Panel) and empiric treatment (Droplet Isolation, Oseltamavir) for the H1N1 swine flu pandemic.[61,62] The viral pandemic dissipated by 2012, with the most prominent orders shifting towards empiric treatment for community acquired pneumonia[63] (azithromycin, levofloxacin) and antibiotic resistant organisms causing health care associated pneumonia[64] (piperacillin-tazobactam).

Table 3 depicts the Rank Biased Overlap (RBO) between 2009 vs. 2012 for each of the most common admission diagnoses. Qualitative patterns reveal more stable ordering patterns (higher RBO) for elective hospital admissions with specific treatment plans and protocols like chemotherapy and symptoms with common initial management (shortness of breath). Greater variability in ordering patterns (lower RBO) is seen for admission diagnoses with dynamically evolving practice patterns and syndromes with less consistent management approaches, such as digestive symptoms and back disorders. Table 4 similarly shows which treatment teams demonstrate more or less practice stability.

Figure 1 and Table 5 report average accuracy metrics for predicting future (2013) clinical order patterns based on models trained on different subsets of historical training data. For all measures, training on separate but concurrent (2013) data consistently yields the best result. This is equivalent to a random train-test split validation, but is not a realistic evaluation of a predictive model that must predict future patterns using only historical data. Training on 12 months of older (2009) historical data performs consistently worse, even compared to training on just 1 month of more recent (2012) data. Expanding the recent (2012) training dataset beyond 12 months yields no further accuracy benefit, and even causes some accuracy decay. Notably, even the "worst" model trained on old (2009) data is substantially more effective than human-authored order sets for anticipating subsequent orders. Figure 2 and Table 6 show a similar set of results for models trained using different "decaying windows" of clinical data relevance.

4 DISCUSSION

The key finding in Table 5 and Figure 1 is that the use of even a small (1 month, ~1.8K patients) but recent (2012) training source is more effective than using a larger (12 month, >10K patients) but older (2009) training source for predicting future (2013) practices. Expanding the recent training source from 1 to 12 months may confer more future predictive power. Continuing to pile on longitudinal data (up to 48 months, 2009–2012) increases the amount of patient data available (~35K patients), but yields no better to slightly worse prediction of future (2013) patterns. This is likely due to the variability of clinical practice patterns over time, making older data less relevant (if not overtly distracting) when predicting future events.

Table 3 and 4 shows a range of rank biased overlap (RBO) values for clinical orders associated with common admission diagnoses and treatment teams, representing the variable stability of clinical order practices. Qualitatively, this supports the general supposition that clinical practices dynamically change over time. Breaking the results down indicate that elective admissions for planned procedures like chemotherapy on the bone marrow transplant service show relatively less variability over time with higher RBO. This could of course be disrupted if future practices shifted in response to different chemotherapy protocols, though the model could still reasonably suggest co-medications that are not enforced through a strict protocol. Diagnoses subject to epidemiologic shifts (e.g., pneumonia) and medical admissions for non-specific symptoms (e.g., digestive) may trigger variable approaches to workup, represented by their lower RBO. Notably, some "non-specific" admission diagnoses like "shortness of breath" exhibit relatively stable patterns.

This indicates that it is not the specificity of the *diagnosis* that confers practice stability but rather the relative consistency with which physicians approach patients with a given symptom or diagnosis.

Table 5 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 and recall at K pay particular attention to the top items that a human user could realistically be expected to review in a decision support order set. As might be expected, predicting 2013 practices is more accurate by all measures when training on concurrent (2013) data than recent (2012) historical data, which in turn is more accurate than using older (2009) historical data. This confirms that prediction performance on typical cross-validated or train-test data splits will likely overestimate performance on future data. The more compelling question answered is whether simply adding more data will yield better results. This is not necessarily the case as we find that 1 month of recent (2012) training data is more effective than 12 months of older (2009) data, and 48 months of data is no better than the most recent 12 months of the same data. While larger datasets are generally expected to improve the power of statistical learning methods, unstable clinical practice patterns can compromise the relevance of data from large expanses of time.

Table 6 reports the accuracy of models trained on all 48 months of historical data available but applying different decaying weights to older data. As in Figure 2, we find that such a decaying weighting scheme can improve future predictions up to and even slightly beyond a simple non-weighted truncation of the most recent 12 months of data. This illustrates an algorithmic approach to continuously stream data into the model while suppressing the loss of accuracy from including older data. Peak accuracy in this case occurs with a decay window about 6 months, reflecting the "mean lifetime" of data relevance. Alternatively, this can be expressed as a half-life of about 4 months for the relevance of historical clinical data towards predicting future practice patterns. These quantitative assessments of dynamic changes in clinical practice areas carry implications for ongoing debates on the appropriate interval for continuing medical education of practicing clinicians,[65,66] the frequency of maintenance and revision required for clinical knowledge content, and the relevance of historical data towards predictive models.

Limitations in our evaluation of clinical practice pattern stability by rank biased overlap of diagnosis-order associations is the presumption that changing patterns reflect 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 changes, such as shifts in diagnosis coding practices from pneumonia to sepsis.[67] Administrative infrastructure changes are expected to occur despite having little semantic difference for clinical decision making, such as hospital orders for Respiratory Virus DFA (direct fluorescent antibody) panels being replaced with Respiratory Virus PCR (polymerase chain reaction) panels. There may also be a substantial shift in patient case mix and characteristics that may not be captured by admission diagnosis stratification, such as admissions for "digestive symptoms" that might represent anything from abdominal pain to diarrhea to vomiting. With all data deriving from a single medical center, significant cultural shifts in practice patterns could also be unduly influenced by a small number of prominent clinicians. This study focuses on inpatient encounters with typically brief durations spanning a matter of days. Thus, while we question the relevance of large longitudinal historical datasets to answer questions of near-term clinical management, such data sources remain important for studying chronic disease, adverse effects, prevention, and epidemiology that may not manifest until after years of observation and for rare conditions with inadequate data quantity in any brief time period to power statistical inquiry.

Even if clinical practice patterns change for “non-clinical” reasons, the overarching caution these results provide for predicting future clinical events based on non-stationary historical data still holds. While prediction models for fixed outcomes may be based more on “fundamental” determinants than time-varying behaviors, prediction and prognosis can still vary substantially with changes in practice. For example, preliminary analysis indicates that applying our described prediction approach towards 30 day hospital mortality will achieve high accuracy (ROC AUC 0.92) using either old (2009) or current (2013) data. Yet, there are clearly cases where prognosis models for certain diseases such as HER2+ breast cancer[68] and chronic Hepatitis C infection[69] dramatically changed in the face of newly developed therapies. Evolving patterns confirm the challenge of manually producing clinical decision support and knowledge guides, as they must be followed by ongoing manual effort to maintain them against practices that may shift within a matter of months. Automated algorithms to learn clinical decision support are thus even more important to not only cover the breadth of medical knowledge, but to automatically adapt to continuous streams of new data. 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. The results above inform such an approach, indicating that prioritizing recent data is more important than accumulating a massive repository of historical data whose interpretation may not even remain internally consistent over time.

This study focuses on the relevance of learned clinical order patterns towards predicting future events, but makes the general presumption that common behaviors reflect “good” decisions that yield better patient outcomes. The “wisdom of the crowd”[70] and Condorcet’s jury theorem[71] posit that aggregated responses from increasing numbers of non-random decision makers will converge towards correctness in the absence of an underlying systematic bias. While such biases are more prone to affect single institutions, it seems unlikely that physicians would be allowed to practice medicine at all if they did not make better than random decisions most of the time. Short of randomized trials, we are evaluating our automated order recommendations against the external standards-of-care established in clinical practice guidelines.[72] With the results of this study however, it is not surprising that practice guidelines themselves must undergo regular revision,[22] 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.

In the absence of a gold standard to define high quality medical decision making, we reference human-authored order sets as the benchmark current standard of care in clinical decision support. As noted in Table 3, even the “worst” association model trained on old

2009 data significantly out-performed manually authored order sets for predicting subsequent clinical orders. Computerized expert systems are unlikely to overtly replace the critical thinking of human clinicians in the foreseeable future, but seeing the successful application of large scale analytics to enable tasks from web document retrieval,[73] commercial product recommendations,[40] to fraud detection,[74] we instead foresee the potential of human clinicians *augmented* with data-driven, point-of-care CDS tools to continually improve care quality and efficiency.

5 CONCLUSION

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) demonstrate more stable practice patterns over time. Small amounts of *recent* training data are more useful than larger amounts of older data towards future predictions. Adding on progressively larger longitudinal training data may actually worsen future prediction accuracy without a weighting scheme to prioritize recent data. For inpatient clinical orders, the half-life of relevance of historical data towards future predictions is about four months.

Acknowledgments

Funding Support: J.H.C supported by the NIH Big Data 2 Knowledge Award Number K01ES026837 through the National Institute of Environmental Health Sciences. 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 Gomathi Krishnan, Lee Ann Yasukawa, Tanya Podchiyska, 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.

References

1. Richardson, WC., Berwick, DM., Bisgard, JC., Bristow, LR., Buck, CR., Cassel, CK., Chassin, MR., Coye, MJ., Detmer, DE., Grossman, JH., James, B., Lawrence, DM., Leape, LL., Levin, A., Robinson-Beale, R., Scherger, JE., Southam, A., Wakefield, M., Warden, GL. Crossing the Quality Chasm: A New Health System for the 21st Century. Institute of Medicine, Committee on Quality of Health Care in America Committee on Quality of Health Care in America; Washington DC: 2001.
2. Lauer MS, Bonds D. Eliminating the "expensive" adjective for clinical trials. Am Heart J. 2014; 167:419–20. DOI: 10.1016/j.ahj.2013.12.003 [PubMed: 24655687]
3. Tricoci P, Allen JM, Kramer JM, Califf RM, Smith SC. Scientific evidence underlying the ACC/AHA clinical practice guidelines. JAMA. 2009; 301:831–41. DOI: 10.1001/jama.2009.205 [PubMed: 19244190]
4. Durack DT. The weight of medical knowledge. N Engl J Med. 1978; 298:773–775. DOI: 10.1097/00006534-197811000-00140 [PubMed: 342963]
5. Alper J, Grossmann C. Health System Leaders Working Toward High-Value Care. 2014
6. Kaushal R, Shojania KG, Bates DW. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. Arch Intern Med. 2003; 163:1409–1416. http://www.ncbi.nlm.nih.gov/pubmed/12824090. [PubMed: 12824090]

7. Overhage J, Tierney W. A randomized trial of "corollary orders" to prevent errors of omission. J Am Med Informatics Assoc. 1997; 4:364–75. [accessed September 25, 2012] http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=61254&tool=pmcentrez&rendertype=abstract.

8. Tierney WM, Overhage JM, Takesue BY, Harris LE, Murray MD, Vargo DL, McDonald CJ. Computerizing Guidelines to Improve Care and Patient Outcomes: The Example of Heart Failure. J Am Med Informatics Assoc. 1995; 2:316–322. DOI: 10.1136/jamia.1995.96073834

9. Chen JH, Fang DZ, Tim Goodnough L, Evans KH, Lee Porter M, Shieh L. Why providers transfuse blood products outside recommended guidelines in spite of integrated electronic best practice alerts. J Hosp Med. 2015; 10:1–7. DOI: 10.1002/jhm.2236

10. Finlay GD, Rothman MJ, Smith Ra. Measuring the modified early warning score and the Rothman index: advantages of utilizing the electronic medical record in an early warning system. J Hosp Med. 2014; 9:116–9. DOI: 10.1002/jhm.2132 [PubMed: 24357519]

11. Kaushal R, Shojania KG, Bates DW. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. Arch Intern Med. 2003; 163:1409–16. DOI: 10.1001/archinte.163.12.1409 [PubMed: 12824090]

12. Cowden, D., Barbacioru, C., Kahwash, E., Saltz, J. Order sets utilization in a clinical order entry system; AMIA Annu Symp Proc. 2003. p. 819http://www.ncbi.nlm.nih.gov/pubmed/14728324http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1479990/pdf/amia2003_0819.pdf

13. Payne TH, Hoey PJ, Nichol P, Lovis C. Preparation and use of preconstructed orders, order sets, and order menus in a computerized provider order entry system. J Am Med Inform Assoc. 2003; 10:322–9. DOI: 10.1197/jamia.M1090 [PubMed: 12668686]

14. Bobb AM, Payne TH, Gross PA. Viewpoint: controversies surrounding use of order sets for clinical decision support in computerized provider order entry. J Am Med Inform Assoc. 2007; 14:41–7. DOI: 10.1197/jamia.M2184 [PubMed: 17068352]

15. Ballard DW, Kim AS, Huang J, Park DK, Kene MV, Chettipally UK, Iskin HR, Hsu J, Vinson DR, Mark DG, Reed ME. 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. DOI: 10.1016/j.annemergmed.2015.07.018

16. Ballesca, Ma, LaGuardia, JC., Lee, PC., Hwang, AM., Park, DK., Gardner, MN., Turk, BJ., Kipnis, P., Escobar, GJ. 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; 9:155–161. DOI: 10.1002/jhm.2149 [PubMed: 24493376]

17. Micek ST, Roubinian N, Heuring T, Bode M, Williams J, Harrison C, Murphy T, Prentice D, Ruoff BE, Kollef MH. Before-after study of a standardized hospital order set for the management of septic shock. Crit Care Med. 2006; 34:2707–13. DOI: 10.1097/01.CCM.0000241151.25426.D7 [PubMed: 16943733]

18. Jacobs BR, Hart KW, Rucker DW. Reduction in Clinical Variance Using Targeted Design Changes in Computerized Provider Order Entry (CPOE) Order Sets: Impact on Hospitalized Children with Acute Asthma Exacerbation. Appl Clin Inform. 2012; 3:52–63. DOI: 10.4338/ACI-2011-01-RA-0002 [PubMed: 23616900]

19. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, Spurr C, Khorasani R, Tanasijevic M, Middleton B. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc. 2003; 10:523–30. DOI: 10.1197/jamia.M1370 [PubMed: 12925543]

20. Goldstein, MK., Hoffman, BB., Coleman, RW., Musen, MA., Tu, SW., Advani, A., Shankar, R., O'Connor, M. [accessed August 29, 2014] 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. 2000. p. 300-4.http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2243943&tool=pmcentrez&rendertype=abstract

21. Hripcsak G, Albers DJ, Perotte a. Exploiting time in electronic health record correlations. J Am Med Informatics Assoc. 2011; 18:i109–i115. DOI: 10.1136/amiajnl-2011-000463

22. Ortiz E, Rhodes S, Morton SC, Eccles MP, Grimshaw JM, Woolf SH. Validity of the Agency for Healthcare Research and Quality Clinical Practice Guidelines. 2001; 286:1461–1467.

23. ONC. Health information technology: standards, implementation specifications, and certification criteria for electronic health record technology, 2014 edition; revisions to the permanent certification program for health information technology. Final rule. Fed Regist. 2012; 77:54163–292. [accessed July 30, 2014] http://www.ncbi.nlm.nih.gov/pubmed/22946139. [PubMed: 22946139]

24. Longhurst, Ca, Harrington, Ra, Shah, NH. A "Green Button" For Using Aggregate Patient Data At The Point Of Care. Health Aff. 2014; 33:1229–1235. DOI: 10.1377/hlthaff.2014.0099

25. Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. N Engl J Med. 2011; 365:1758–9. DOI: 10.1056/NEJMp1108726 [PubMed: 22047518]

26. Smith, M., Saunders, R., Stuckhardt, L., McGinnis, JM. Best care at lower cost: the path to continuously learning health care in America. Institute of Medicine, Committee on the Learning Health Care System in America; Washington DC: 2012.

27. Krumholz HM. Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. Health Aff. 2014; 33:1163–1170. DOI: 10.1377/hlthaff.2014.0053

28. de Lissovoy G. Big data meets the electronic medical record: a commentary on "identifying patients at increased risk for unplanned readmission". Med Care. 2013; 51:759–60. DOI: 10.1097/MLR.0b013e3182a67209 [PubMed: 23942217]

29. Sittig DF, Wright A, Osheroff Ja, Middleton B, Teich JM, Ash JS, Campbell E, Bates DW. Grand challenges in clinical decision support. J Biomed Inform. 2008; 41:387–92. DOI: 10.1016/j.jbi.2007.09.003 [PubMed: 18029232]

30. Doddi S, Marathe a, Ravi SS, Torney DC. Discovery of association rules in medical data. Med Inform Internet Med. 2001; 26:25–33. http://www.ncbi.nlm.nih.gov/pubmed/11583406. [PubMed: 11583406]

31. Klann J, Schadow G, Downs SM. A method to compute treatment suggestions from local order entry data. AMIA Annu Symp Proc. 2010; 2010:387–91. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3041352&tool=pmcentrez&rendertype=abstract. [PubMed: 21347006]

32. Klann J, Schadow G, McCoy JM. A recommendation algorithm for automating corollary order generation. AMIA Annu Symp Proc. 2009; 2009:333–7. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2815486&tool=pmcentrez&rendertype=abstract. [PubMed: 20351875]

33. Wright A, Sittig DF. Automated development of order sets and corollary orders by data mining in an ambulatory computerized physician order entry system. AMIA Annu Symp Proc. 2006; 2006:819–823. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17238455.

34. Zhang Y, Padman R, Levin JE. Paving the COWpath: data-driven design of pediatric order sets. J Am Med Inform Assoc. 2014; 21:e304–e311. DOI: 10.1136/amiajnl-2013-002316 [PubMed: 24674844]

35. Klann JG, Szolovits P, Downs SM, Schadow G. Decision support from local data: creating adaptive order menus from past clinician behavior. J Biomed Inform. 2014; 48:84–93. DOI: 10.1016/j.jbi.2013.12.005 [PubMed: 24355978]

36. Wright AP, Wright AT, McCoy AB, Sittig DF. The use of sequential pattern mining to predict next prescribed medications. J Biomed Inform. 2014; 53:73–80. DOI: 10.1016/j.jbi.2014.09.003 [PubMed: 25236952]

37. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. J Biomed Inform. 2010; 43:891–901. DOI: 10.1016/j.jbi.2010.09.009 [PubMed: 20884377]

38. 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; 2013:34–8. [accessed October 27, 2014] http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3845792&tool=pmcentrez&rendertype=abstract.

39. 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; 23:339–348. DOI: 10.1093/jamia/ocv091

40. Linden G, Smith B, York J. Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Comput. 2003; 7:76–80. DOI: 10.1109/MIC.2003.1167344

41. Moskovitch R, Choi H, Hripcsak G, Tatonetti N. Prognosis of Clinical Outcomes with Temporal Patterns and Experiences with One Class Feature Selection. IEEE/ACM Trans Comput Biol Bioinforma. 2016; 5963:1–1. DOI: 10.1109/TCBB.2016.2591539

42. Coiera E, Wang Y, Magrabi F, Concha OP, Gallego B, Runciman W. Predicting the cumulative risk of death during hospitalization by modeling weekend, weekday and diurnal mortality risks. BMC Health Serv Res. 2014; 14:226.doi: 10.1186/1472-6963-14-226 [PubMed: 24886152]

43. Gallego B, Magrabi F, Concha OP, Wang Y, Coiera E. Insights into temporal patterns of hospital patient safety from routinely collected electronic data. Heal Inf Sci Syst. 2015; 3:S2.doi: 10.1186/2047-2501-3-S1-S2

44. Lemeshow S, Le Gall JR. Modeling the severity of illness of ICU patients. A systems update. JAMA. 1994; 272:1049–55. [accessed July 31, 2014] http://www.ncbi.nlm.nih.gov/pubmed/8089888. [PubMed: 8089888]

45. Kansagara D. Risk Prediction Models for Hospital Readmission<subtitle>A Systematic Review</subtitle>. JAMA J Am Med Assoc. 2011; 306:1688.doi: 10.1001/jama.2011.1515

46. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med. 2015; 7:299ra122.doi: 10.1126/scitranslmed.aab3719

47. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Informatics Assoc. 2016; 2016:ocw042.doi: 10.1093/jamia/ocw042

48. Bergmeir C, Hyndman RJ, Koo B. A Note on the Validity of Cross-Validation for Evaluating Time Series Prediction. 2015

49. Lowe HJ, Ferris Ta, Hernandez PM, Weber SC. STRIDE--An integrated standards-based translational research informatics platform. AMIA Annu Symp Proc. 2009; 2009:391–5. [PubMed: 20351886]

50. Hernandez P, Podchiyska T, Weber S, Ferris T, Lowe H. Automated mapping of pharmacy orders from two electronic health record systems to RxNorm within the STRIDE clinical data warehouse. AMIA Annu Symp Proc. 2009; 2009:244–8. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2815471&tool=pmcentrez&rendertype=abstract. [PubMed: 20351858]

51. Wright A, Bates DW. Distribution of Problems, Medications and Lab Results in Electronic Health Records: The Pareto Principle at Work. Appl Clin Inform. 2010; 1:32–37. DOI: 10.4338/ACI-2009-12-RA-0023 [PubMed: 21991298]

52. 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; 2014:206–10. http://www.ncbi.nlm.nih.gov/pubmed/25717414.

53. Finlayson SG, Lependu P, Shah NH. Building the graph of medicine from millions of clinical narratives. Sci Data. 2014; 1:140032.doi: 10.1038/sdata.2014.32 [PubMed: 25977789]

54. Manrai AK, Bhatia G, Strymish J, Kohane IS, Jain SH. Medicine’s uncomfortable relationship with math: calculating positive predictive value. JAMA Intern Med. 2014; 174:991–3. DOI: 10.1001/jamainternmed.2014.1059 [PubMed: 24756486]

55. Kendall MG. A New Measure of Rank Correlation. Biometrika. 1938; 30:81–93. DOI: 10.2307/2332226

56. Webber W, Moffat A, Zobel J. A similarity measure for indefinite rankings. ACM Trans Inf Syst. 2010; 28:1–38. DOI: 10.1145/1852102.1852106

57. Agrawal, R. [accessed February 3, 2015] Comparing Ranked List. 2013. https://ragrawal.wordpress.com/2013/01/18/comparing-ranked-list/

58. Chen JH, Goldstein MK, Asch SM, Mackey L, Altman RB. Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. J Am Med Informatics Assoc. 2016; :ocw136.doi: 10.1093/jamia/ocw136

59. Jones, E., Oliphant, T., Peterson, P., Al, E. [accessed March 1, 2015] SciPy: Open source scientific tools for Python. n.d. http://www.scipy.org

60. Leskovec, J., Rajaraman, A., Ullman, JD. Min Massive Datasets. 2014. Mining Data Streams; p. 131-162.

61. Kerr JR. Swine influenza. J Clin Pathol. 2009; 62:577–578. DOI: 10.1136/jcp.2009.067710 [PubMed: 19433408]

62. 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. 2010. p. 1-32.http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:WHO,2010#7

63. Mandell, La, Wunderink, RG., Anzueto, A., Bartlett, JG., Campbell, GD., Dean, NC., Dowell, SF., File, TM., Musher, DM., Niederman, MS., Torres, A., Whitney, CG. Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. Clin Infect Dis. 2007; 44(Suppl 2):S27–72. DOI: 10.1086/511159 [PubMed: 17278083]

64. Focaccia R, Gomes Da Conceicao OJ. Guidelines for the management of adults with hospital-acquired, ventilator-associated, and healthcare-associated pneumonia. Am J Respir Crit Care Med. 2005; 171:388–416. DOI: 10.1164/rccm.200405-644ST [PubMed: 15699079]

65. Teirstein P. Boarded to Death—Why Maintenance of Certification Is Bad for Doctors and Patients. N Engl J Med. 2015; 372:106–8. DOI: 10.1056/NEJMp1407422 [PubMed: 25564895]

66. Irons MB, Nora LM. Maintenance of Certification 2.0—Strong Start, Continued Evolution. N Engl J Med. 2015; 372:104–106. DOI: 10.1056/NEJMp1409923 [PubMed: 25564894]

67. Rhee C, Gohil S, Klompas M. Regulatory mandates for sepsis care--reasons for caution. N Engl J Med. 2014; 370:1673–1676. DOI: 10.1056/NEJMp1400276 [PubMed: 24738642]

68. Wang H, Zhang C, Zhang J, Kong L, Zhu H, Yu J. The prognosis analysis of different metastasis pattern in patients with different breast cancer subtypes : a SEER based study. 2016

69. Bhatia HK, Singh H, Grewal N, Natt NK. Sofosbuvir: A novel treatment option for chronic hepatitis C infection. J Pharmacol Pharmacother. 2014; 5:278–284. DOI: 10.4103/0976-500X.142464 [PubMed: 25422576]

70. Surowiecki, J. The wisdom of crowds: why the many are smarter than the few and how collective wisom shapes business, economies, societies, and nations. Doubleday; New York: 2004. http://books.google.com/books?id=bA0c4aYTD6gC&pgis=1

71. Austen-Smith D, Banks JS. Information Aggregation, Rationality, and the Condorcet Jury Theorem. Am Polit Sci Rev. 1996; 90:34–45. DOI: 10.2307/2082796

72. Chen JH, Altman RB. Data-Mining Electronic Medical Records for Clinical Order Recommendations: Wisdom of the Crowd or Tyranny of the Mob? AMIA Jt Summits Transl Sci Proc AMIA Summit Transl Sci. 2015; 2015:435–9. http://www.ncbi.nlm.nih.gov/pubmed/26306281.

73. Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine BT - Computer Networks and ISDN Systems. 1998; 30:107–117. DOI: 10.1016/S0169-7552(98)00110-X

74. Excell D. Bayesian inference - The future of online fraud protection. Comput Fraud Secur. 2012; 2012:8–11. DOI: 10.1016/S1361-3723(12)70018-0

Highlights

- *Recent* data beats *more* data when predicting the future
- Clinical order patterns algorithmically learned from real-world data predict clinical practices more effectively than manually authored decision support
- The "half-life" of clinical data relevance towards predicting future practices is about four months

What was already known?

- Clinical decision support in the form of order sets, alerts, and risk scores improve clinical practice, but...
- Manual authoring limits the reach of clinical decision support, especially when clinical best practices become obsolete and require ongoing revision
- The advent of pervasive electronic medical record data create the opportunity for large-scale data-driven clinical predictions and decision support

What this study added to our knowledge?

- Clinical order patterns algorithmically learned from real-world data is significantly more effective than manually authored decision support (order sets) towards anticipating real clinical practices
- Prioritizing small amounts of recent data is more effective than using larger amounts of older data towards predicting future clinical decisions

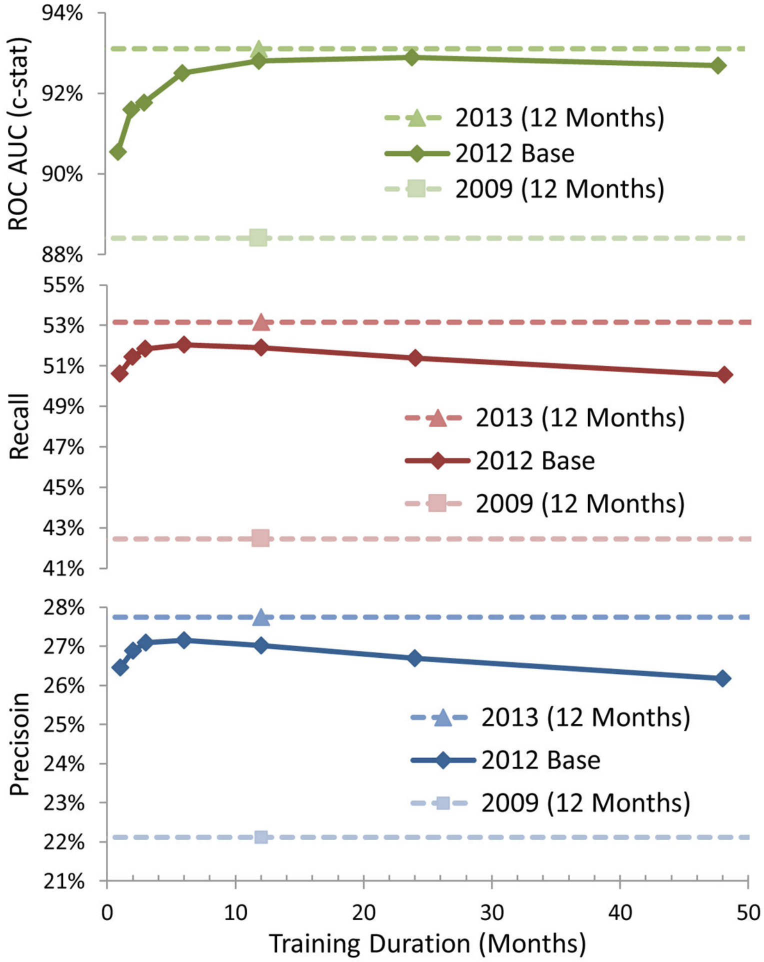

Figure 1.

Accuracy predicting 2013 admission orders when using different subsets of historical training data. Training on separate but concurrent (2013) data (top horizontal bar) is equivalent to a random train-test split validation. Training on 12 months of older (2009) historical data (bottom horizontal bar) performs consistently worse. Expanding recent (2012) training dataset from 1 up to 48 months varies future prediction accuracy.

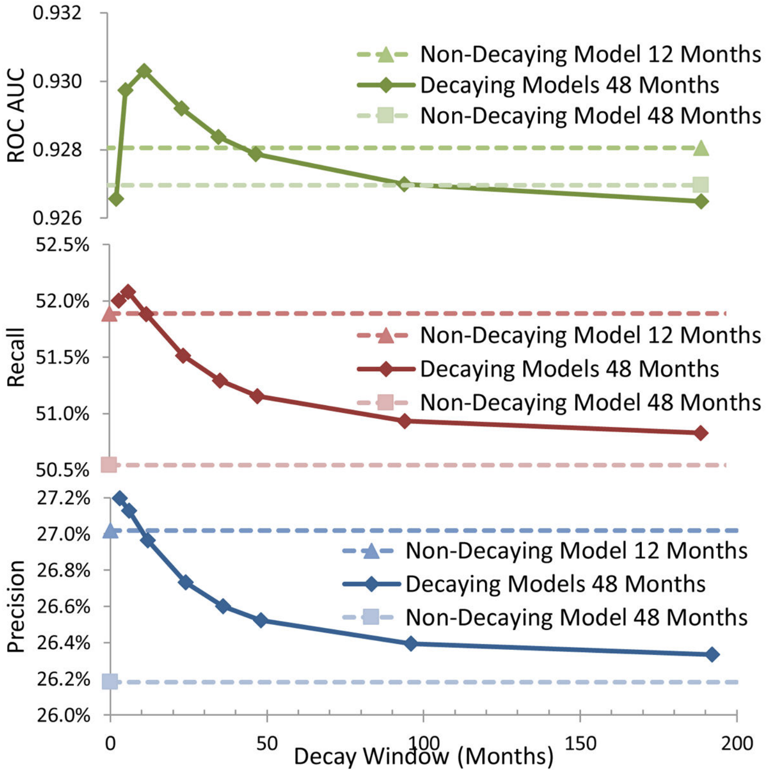

Figure 2.

Accuracy predicting inpatient orders in 2013 when using historical training data from 2012 and before. Using only the most recent 12 months of data (top horizontal bar) yields better future predictions than using 48 months of prior data (bottom horizontal bar). Using all 48 months of prior data can yield progressively better future predictions by applying a decaying weighting scheme that discounts older data in favor of recent data.

Table 1

Example non-zero counts per ICD9 admission diagnosis from 2008–2014. Noise (extra counts) have been added to avoid potentially identifying data bins with counts <10. Detailed five digit ICD9 codes often lead to sparse elements, such as only a handful of admissions coded as 787.24. To compress the hierarchy, instances of five digit codes (e.g., 787.24) were also counted as the respective four digit code (e.g., 787.2), which were in turn also counted as the three digit code (e.g., 787). Thus, the aggregated 787.2 admission diagnosis code accounts for direct codes for 787.2, as well as all instances of 787.2x sub-codes. Likewise, the 787 admission diagnosis code accounts for all 787.x and 787.xx sub-codes.

Raw Count	Aggregate Count	ICD9	Description
0	1934	787	Symptoms involving digestive system
0	1111	787.0	Nausea and vomiting
872	872	787.01	Nausea with vomiting
100	100	787.02	Nausea alone
125	125	787.03	Vomiting alone
14	14	787.04	Bilious emesis
0	259	787.2	Dysphagia
215	215	787.20	Dysphagia, unspecified
13	13	787.22	Dysphagia, oropharyngeal phase
11	11	787.24	Dysphagia, pharyngoesophageal phase
20	20	787.29	Other dysphagia
83	83	787.3	Flatulence, eructation, and gas pain
3	17	787.6	Incontinence of feces
14	14	787.60	Full incontinence of feces
0	464	787.9	Other symptoms involving digestive system
450	450	787.91	Diarrhea
14	14	787.99	Other symptoms involving digestive system

Table 2

Top clinical order associations for admission diagnosis of “Pneumonia” (ICD9: 486) in 2009 vs. 2012. Based on orders occurring within one day of the admission diagnosis and ranked by P-value (chi-square with Yates’ correction). At each rank *k*, the intersection of the top *k* items from each list defines the “Overlap at Rank Depth.” The ratio of overlap to rank yields a “Fractional Overlap.” For the full list of 2,030 candidate clinical orders, averaging the Fractional Overlap column defines an Average Overlap score. Averaging with a geometric weighting scheme to emphasize the importance of top suggestions and ensure numerical convergence yields a Rank Biased Overlap (RBO) score. Rank Biased Overlap (RBO) = 0.49, indicating a moderate shift in the item rankings between the two lists.

Top Orders (2009)	Overlap at Rank Depth	Rank	Fractional Overlap	Top Orders (2012)
Respiratory Virus Panel	0	1	0.00	Blood Culture (2x Aerobic)
Blood Culture (2x Aerobic)	1	2	0.50	Blood Culture (Aerobic+ Anaerobic)
Blood Culture (Aerobic+ Anaerobic)	2	3	0.67	Azithromycin (IV)
PoC Venous Blood Gas + Lactate	2	4	0.50	Levofloxacin (IV)
PoC Venous Blood Panel	3	5	0.60	PoC Venous Blood Gas + Lactate
Levofloxacin (IV)	4	6	0.67	PoC Troponin I
Droplet Isolation	4	7	0.57	Legionella Antigen, Urine
Oseltamivir (Oral)	4	8	0.50	Consult to Medicine
Respiratory - Nebulizer Treatment	4	9	0.44	Xray Chest 2 View
Respiratory Culture	4	10	0.40	Piperacillin-Tazobactam (IV)
…	…	…	…	…

Table 3

Most common admission diagnoses in 2012 and Rank Biased Overlap (RBO) of associated orders from 2009 vs. 2012 that reflect the relative stability in top associated orders over time.

Rank Biased Overlap 2009 vs. 2012	Admissions in 2012	Admission Diagnosis (ICD9)
0.68	257	Encounter for chemotherapy/immunotherapy (V58.1)
0.68	267	Encounter for procedures and aftercare (V58)
0.66	239	Encounter for antineoplastic chemotherapy (V58.11)
0.60	845	Respiratory and chest symptoms (786)
0.59	351	Shortness of breath (786.05)
0.59	938	General symptoms (780)
0.59	388	Specific procedure complications (996)
0.58	442	Dyspnea (786.0)
0.57	662	Osteoarthrosis (715)
0.57	202	Other general symptoms (780.9)
0.57	196	Altered mental status (780.97)
0.55	365	Osteoarthrosis, localized (715.3)
0.54	307	Fever and temperature dysregulation (780.6)
0.54	283	Fever, unspecified (780.60)
0.54	191	Gastrointestinal hemorrhage (578)
0.53	311	Chest pain (786.5)
0.51	458	Abdominal pain (789.0)
0.51	245	Chest pain, unspecified (786.50)
0.48	187	Osteoarthrosis, lower leg (715.36)
0.48	532	Abdomen and pelvis symptoms (789)
0.47	228	Osteoarthrosis, unspecified (715.9)
0.44	214	Abdominal pain, unspecified site (789.00)
0.44	178	Cardiac dysrhythmias (427)
0.42	261	Back disorders (724)
0.33	240	Symptoms involving digestive system (787)

Table 4

Most common medical treatment teams assigned to Primary or Consulting roles in patient care, with the respective number of assignments in 2012. Rank Biased Overlap (RBO) of associated orders from 2009 vs. 2012 reflect the relative stability in top associated orders that occur within 24 hours of the given team assignment. Abbreviations: ICU = Intensive Care Unit.

Rank Biased Overlap 2009 vs. 2012	Assignments in 2012	Treatment Team (Primary vs. Consulting)
0.62	301	Bone Marrow Transplant (Primary)
0.62	969	Medical ICU (Primary)
0.61	475	Cardiovascular ICU (Primary)
0.59	221	Neurology (Primary)
0.56	455	Neurology ICU (Consulting)
0.55	413	Nephrology (Consulting)
0.54	477	Surgical ICU (Primary)
0.49	430	Oncology (Primary)
0.46	259	Neurology Stroke (Consulting)
0.44	285	Medical ICU (Consulting)
0.44	241	Hematology (Primary)
0.36	202	Nephrology Transplant (Consulting)
0.36	94	Neurology Epilepsy Monitor Unit (Primary)
0.34	328	Neurology General (Consulting)
0.32	86	Kidney Pancreas Transplant (Primary)
0.31	619	Acute Care Surgery Trauma (Primary)
0.29	87	Cystic Fibrosis Adult (Primary)
0.27	150	Hematology (Consulting)
0.25	531	Acute Care Surgery Trauma (Consulting)
0.25	66	Liver Transplant (Primary)
0.22	1811	Medicine University (Primary)
0.22	191	General Thoracic (Primary)
0.22	1006	Private Medicine (Primary)
0.20	342	Cardiac Adult (Primary)
0.20	216	Vascular Surgery (Consulting)

Table 5

Accuracy for predicting inpatient orders when using different subsets of historical training data. For 4,820 hospital admissions in 2013, clinical data up to the use of a pre-authored hospital order set was used to query a previously trained association matrix for a score-ranked list of clinical orders. Order lists were score-ranked by PPV (positive predictive value ~ post-test probability) relative to the query items to identify orders most *likely* to subsequently occur and compared against the subsequent 24 hours of clinical orders that actually occurred in the 2013 admission. Full list ranking is evaluated by the area under the receiver operating characteristic curve (ROC-AUC), while precision and recall at K items evaluates only the top associations. Compared to the 2013 and 2009 results, all other results differed by two-tailed *t*-tests with $P<10^{-10}$.

Training Base Year	Training Duration (Months)	Patients	Clinical Item Instances	Precision	Recall	ROC AUC
2013	12	11,278	3,346,201	27.7%	53.2%	0.931
2012	1	1,825	288,199	26.5%	50.6%	0.906
2012	2	3,079	557,502	26.9%	51.4%	0.916
2012	3	4,243	832,515	27.1%	51.8%	0.918
2012	6	7,284	1,686,412	27.2%	52.0%	0.925
2012	12	12,503	3,326,376	27.0%	51.9%	0.928
2012	24	21,901	6,543,654	26.7%	51.4%	0.929
2012	48	34,812	11,522,166	26.2%	50.5%	0.927
2009	12	10,727	1,838,637	22.1%	42.5%	0.884
Pre-Authored Order Sets				15.9%	35.1%	0.812

Table 6

Accuracy for predicting inpatient orders in 2013 when using different exponential decaying data parameters on 48 months of historical training data. Peak accuracy is observed with a decay window about 6 months.

Decay Window (Months)	Decay Interval (Weeks)	Precision	Recall	ROC AUC (c-stat)
3	4	27.2%	52.0%	0.927
6	4	27.1%	52.1%	0.930
12	4	27.0%	51.9%	0.930
24	4	26.7%	51.5%	0.929
36	4	26.6%	51.3%	0.928
48	4	26.5%	51.2%	0.928
96	4	26.4%	50.9%	0.927
192	4	26.3%	50.8%	0.926
36	1	26.6%	51.2%	0.927
36	2	26.6%	51.2%	0.927
36	4	26.6%	51.3%	0.928
36	8	26.7%	51.4%	0.930