


Research and Applications

OrderRex clinical user testing: a randomized trial of recommender system decision support on simulated cases

Andre Kumar ¹ Rachael C. Aikens,^{2,3} Jason Hom,¹ Lisa Shieh,¹ Jonathan Chiang,⁴ David Morales,⁵ Divya Saini,⁵ Mark Musen,⁴ Michael Baiocchi,⁶ Russ Altman,⁷ Mary K Goldstein,^{8,9} Steven Asch,^{10,11} and Jonathan H. Chen^{1,4}

¹Division of Hospital Medicine, Department of Medicine, Stanford University, Stanford, California, USA, ²Program in Biomedical Informatics, Stanford University, Stanford, California, USA, ³Department of Statistics, Stanford University, Stanford, California, USA, ⁴Department of Medicine, Center for Biomedical Informatics Research, Stanford University, Stanford, California, USA, ⁵Department of Computer Science, Stanford University, Stanford, California, USA, ⁶Department of Epidemiology and Public Health, Stanford University, Stanford, California, USA, ⁷Departments of Bioengineering, Genetics, Medicine, and Biomedical Data Science, Stanford University, Stanford, California, USA, ⁸Geriatrics Research Education and Clinical Center, Veteran Affairs Palo Alto Health Care System, Palo Alto, California, USA, ⁹Primary Care and Outcomes Research (PCOR), Department of Medicine, Stanford University, Stanford, California, USA, ¹⁰Primary Care and Population Health, Department of Medicine, Stanford University, Stanford, California, USA, and ¹¹Center for Innovation to Implementation, Veteran Affairs Palo Alto Health Care System, Palo Alto, California, USA

Corresponding Author: Jonathan Chen, MD, PhD, Stanford University School of Medicine, 1265 Welch Rd, Medical School Office Building X213, Stanford, CA 94305, USA (jonc101@stanford.edu)

Received 27 March 2020; Revised 13 July 2020; Editorial Decision 16 July 2020; Accepted 25 July 2020

ABSTRACT

Objective: To assess usability and usefulness of a machine learning-based order recommender system applied to simulated clinical cases.

Materials and Methods: 43 physicians entered orders for 5 simulated clinical cases using a clinical order entry interface with or without access to a previously developed automated order recommender system. Cases were randomly allocated to the recommender system in a 3:2 ratio. A panel of clinicians scored whether the orders placed were clinically appropriate. Our primary outcome included the difference in clinical appropriateness scores. Secondary outcomes included total number of orders, case time, and survey responses.

Results: Clinical appropriateness scores per order were comparable for cases randomized to the order recommender system (mean difference -0.11 order per score, 95% CI: [-0.41, 0.20]). Physicians using the recommender placed more orders (median 16 vs 15 orders, incidence rate ratio 1.09, 95%CI: [1.01-1.17]). Case times were comparable with the recommender system. Order suggestions generated from the recommender system were more likely to match physician needs than standard manual search options. Physicians used recommender suggestions in 98% of available cases. Approximately 95% of participants agreed the system would be useful for their workflows.

Discussion: User testing with a simulated electronic medical record interface can assess the value of machine learning and clinical decision support tools for clinician usability and acceptance before live deployments.

Conclusions: Clinicians can use and accept machine learned clinical order recommendations integrated into an electronic order entry interface in a simulated setting. The clinical appropriateness of orders entered was comparable even when supported by automated recommendations.

Key words: informatics, clinical care, clinical decision support, recommender systems, human computer interaction, usability testing, collaborative filtering, order sets, electronic medical records, clinical provider order entry

INTRODUCTION

Physician compliance with evidence-based care often falls short, with overall compliance with clinical guideline recommendations ranging from 20 to 80%.¹ Such variability may compromise care quality and cost effectiveness, especially when knowledge is inconsistently applied.² The meaningful use era of electronic health records (EHRs)³ creates the opportunity for data-driven clinical decision support (CDS) that utilizes the collective expertise of many practitioners in a learning health system.^{4–8} It may additionally facilitate the acquisition of medical knowledge by enabling clinicians to adopt evolving evidence-based practice patterns.⁹ Tools such as order sets already reinforce consistency and compliance with best practices,^{10,11} but maintainability is limited in scale by a top-down, knowledge-based approach requiring the manual effort of human experts.¹² Moreover, the intended vs actual usage of EHR order sets may not align with physician workflows,¹³ and it may impede physicians from learning appropriate alternatives toward patient care.¹⁴ A key challenge to fulfill a future vision for clinical decision support^{15,16} is the automatic production and delivery of content from the bottom-up by data-mining clinical data sources.¹⁷

Most prior studies in automated development of clinical decision support content have been strictly offline analytical evaluations,^{17–24} with few studies assessing the response of human clinicians to such recommender tools and their ordering patterns. More broadly, the majority of physicians have significant distrust or negative attitudes toward the EHR,^{25–27} which may affect how well these tools could be adopted. As with many machine-learning models designed to support clinical decision-making, it is unknown if physicians will actually accept such suggestions into their clinical workflow.

Previously, we developed an automated order recommender system by data-mining our hospital's EHR data.²² The results of this approach align with established standards of care^{17,28,29} and are predictive of real physician behavior and patient outcomes.²² Our underlying vision is to seamlessly integrate a system into clinical order entry workflows that automatically infers the relevant clinical context based on data already in the EHR and provides actionable decision support in the form of clinical order suggestions, analogous to Netflix or Amazon.com's "customers who bought A also bought B" system.^{30,31} It is unknown if these suggested orders would be accepted by clinicians or affect quality of care.

This study seeks to address these issues by examining physicians' behaviors while interacting with a clinical provider order entry (CPOE) interface that simulates an EHR for hospital clinical scenarios. We specifically examine whether the automated order recommender system impacted the number of clinically inappropriate/appropriate orders placed during the simulated cases. We further evaluated physician ordering patterns, user experience metrics, and survey responses when an automated order recommender system was added to standard functionality.

OBJECTIVE

To determine how clinicians interact with an automated clinical order recommender system for electronic order entry for simulated

clinical cases and whether such recommendations impact the clinical appropriateness of the orders being placed or physician workflow.

MATERIALS AND METHODS

Participants and setting

This study was conducted at a single academic institution from 10/2018–12/2019. We recruited physicians (n=43) with experience caring for medical inpatients within the past year using local mailing listservs. Participants included medical residents (trainees who have a medical license but still require oversight) and attending physicians. The study was approved by the Stanford University Institutional Review Board.

Study design

Participants were offered a \$195 incentive payment for a 1-hour usability testing session in a closed office setting where they were exposed to a series of 5 clinical cases that simulate common inpatient medical problems (see *Cases & Grading* below) on a digital interface that simulated their institution's EHR (Figures 1–3). Upon recruitment to the study, a researcher guided participants through 2 demonstration cases (diabetic ketoacidosis and chest pain) to illustrate basic functions of the digital interface (data review, order entry, order sets).

All physician participants were subsequently assigned each of the 5 cases (Table 1) in random order for a within-subjects design. Three of their 5 cases were randomly assigned to have an automated order recommender system that provided order suggestions (Figure 3) vs no order recommender system in their remaining 2 cases (Figure 2). Based on previous pilot testing, we found that users could not consistently complete more than 5 simulated cases during the scheduled 1-hour testing sessions.³⁷ The unbalanced 3:2 treatment assignment was therefore selected to acquire more usability feedback on the recommender interface. The study participants were unblinded to their treatment allocation as this was infeasible from a user interface study context. Conventional clinical order entry options including order-set checklists and manual search of individual orders were available in all cases, making usage of the automated order recommender system completely optional. Participant activity was recorded through screen capture, audio, and user interface tracking software. Following the case series, all participants filled out a survey on their experiences with the system and their receptiveness to the automated order recommender.

Outcomes

The primary outcome was the clinical appropriateness of orders placed (mean score per order) in the simulated cases. Clinical appropriateness was determined by a panel of clinicians who assigned a score to each order for each simulated case (see *Cases & Grading* below). Secondary outcomes included 4 ordering outcomes (the total score of all orders, number of orders, number of positively scoring orders, and number of nonpositively scoring orders) and 2 user-experience outcomes (number

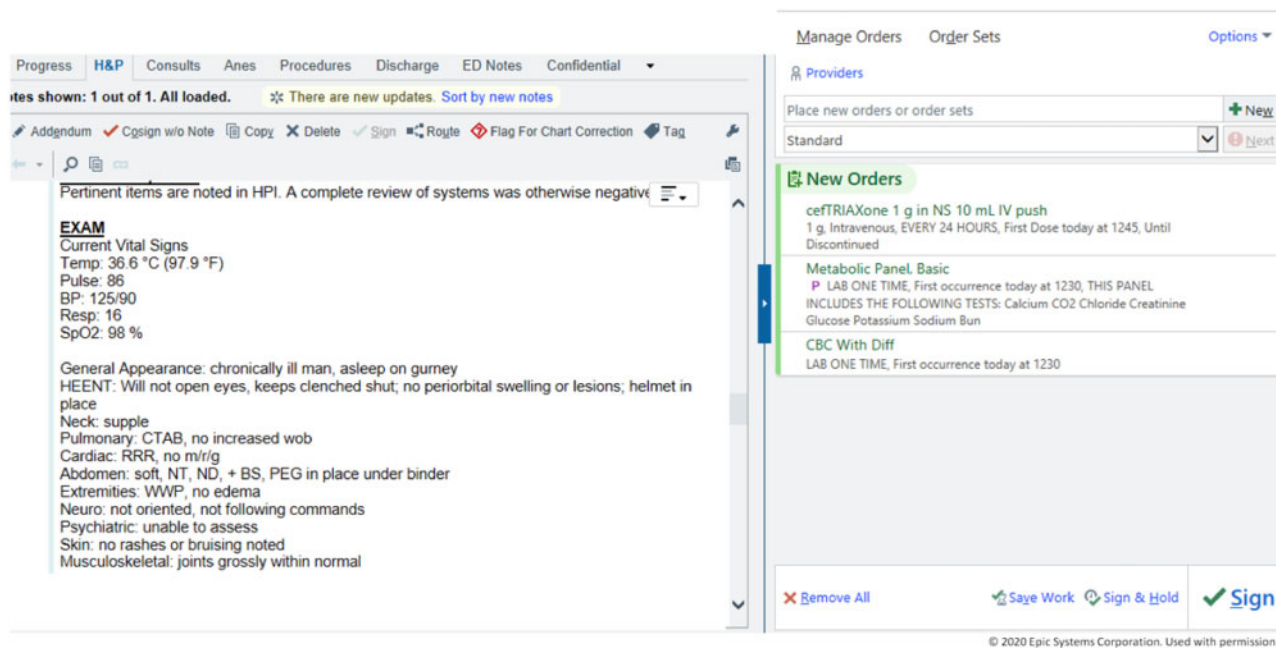


Figure 1. Screenshot from the Epic electronic medical record used in our local hospital.

Note: There is a clinical note window on the left and a clinical order entry interface on the right, including a search box for individual orders and order sets, as well as a running list of new orders placed. All participants in the study were already familiar with this interface from their prior clinical practice.

of clicks, time to complete the simulated cases). Planned ancillary analyses were done to summarize survey results and calculate precision and recall of the ordering systems.

Automated order recommender development

As described previously,²² we extracted deidentified structured data for all inpatient hospitalizations from the 2009–2014 STRIDE clinical data warehouse.³⁸ The data cover > 74K patients with > 11M instances of > 27K items (medication, laboratory, imaging, and nursing orders as well as lab results and diagnosis codes). We built a clinical collaborative filtering (recommender) system based on this data, modeled on commercial product recommender algorithms^{30,31} using item co-occurrence statistics. While the algorithm is specifically designed to adapt to continuous streams of clinical data, to maintain consistency within this study, the recommender system content was kept fixed after training with data up to 2014.

We built a simulated computerized provider order entry (CPOE) interface with open technologies including PostgreSQL, Python, Apache HTTP, and HTML/JavaScript. This CPOE was modeled after the EHR currently used in our hospital (Figures 1 and 2). Our unique addition is an automated recommender (Figure 3), analogous to a “customers who bought this item also bought this...” service that anticipates other clinical orders that are likely to be relevant based on similar prior cases in prior EHRs.

Cases & grading

A panel of board-certified internal medicine physicians (AK, JH, LS, and JHC) developed 5 simulated clinical cases of common inpatient medical problems: unstable atrial fibrillation, neutropenic fever, variceal gastrointestinal hemorrhage, bacterial meningitis, and acute pulmonary embolism (Table 1). Each participant was exposed to the clinical interface (Appendix), which included the patient’s history and physical examination. Depending on the interventions ordered,

the case would progress across several decisional nodes (Table 1). For example, if a participant ordered a lumbar puncture and antibiotics for bacterial meningitis, the case would progress toward a different node (patient improvement). In contrast, if antibiotics were not ordered, the patient would deteriorate (updated vitals and clinical notes would appear in this node). Diagnostic test results are only visible if respective orders are entered. With each entered order, the automated order recommender lists would continuously update based on the accumulating patient information.

To determine clinical appropriateness for all orders, the case designers reviewed options via the Delphi method.³⁹ Each physician panelist (AK, JH, LS, and JHC) independently reviewed all orders placed for the case. Cases were classified according to their state (initial, subsequent, or resolution) and orders were considered in the context of each state. Each grader assigned an individual score to an order on a –10 (very inappropriate) to +10 (very appropriate) scale based on 1-step intervals. The Appendix includes an extended description of how scoring was considered. The initial independent review had an intraclass correlation coefficient (ICC) of 0.51 (95% CI: [0.47–0.53]) for all scored orders on a –10 – +10 scale.³⁹

Following independent scoring, the reviewers met as a panel in multiple rounds to review their independent assessments and deliberated to assign a consensus score.³⁹ Appropriate research studies and clinical guidelines were considered when assigning a consensus score (Table 1). In instances where the panel could not reach a final consensus, no score was assigned.

When considering the scoring outcomes, the total score assigned for a simulated case was the sum of scores for each order placed during the case. The average score per order was the total score divided by the total number of graded orders entered per case. Only unique orders were counted for this graded total, discounting repeats and clinically redundant orders (eg, ordering a complete blood count and a complete blood count with automated differential). Repeat or clinically redundant orders were counted towards the “Total Orders” secondary outcome.

Table 1. Summary description of simulated cases tested

Presenting Symptom (ICD-10) / Diagnosis	Case Summary	Important Decisional Nodes	Most Common Orders (% Frequency)
Fever (453.3) Chemotherapy Induced Neutropenic Fever	32-year-old patient with diffuse large B-cell lymphoma presenting with fevers and rigors after receiving chemotherapy (R-CHOP) 10 days prior. Key Clinical Findings: hypotension, lactic acidosis, severe neutropenia	<i>Patient improves</i> with isotonic fluid resuscitation, 4th generation cephalosporin or piperacillin-tazobactam ³² <i>Patient deteriorates</i> without fluid resuscitation or appropriate antimicrobial coverage	Sodium Chloride IV (98) Comprehensive Metabolic Panel (98) Blood Cultures (93) CBC with Differential (93) Cefepime, IV (88) Chest X-ray (81) Urine Culture (77)
Headache (R55) Bacterial Meningitis	25-year-old previously healthy patient presenting with fever, headache, neck stiffness, and photophobia. Key Clinical Findings: fever, nuchal rigidity, absence of rashes	<i>Patient improves</i> with immediate lumbar puncture, IV ceftriaxone + vancomycin ³³ <i>Patient deteriorates</i> without immediate lumbar puncture and antimicrobials (eg, if the clinician waits 45 minutes order and review CT Head results before ordering a lumbar puncture and antibiotics)	CBC with Differential (95) Ceftriaxone, IV (93) CSF Culture and Gram Stain (93) Glucose, CSF (91) Protein, CSF (88) Cell Count, CSF (84) Comprehensive Metabolic Panel (84) Sodium Chloride IV (83) ECG 12-Lead (91) CBC with Differential (88) Comprehensive Metabolic Panel (77)
Dyspnea (R06.00) Acute Pulmonary Embolism and presumptive lung cancer	70-year-old with a past medical history including systolic heart failure, COPD, and smoking presenting with worsening dyspnea following a vacation to Hawaii Key Clinical Findings: Hypoxia (81% oxygen saturation), tachycardia, absence of jugular distension, minimal wheezes.	<i>Patient improves</i> with oxygenation + therapeutic anticoagulation (heparin, low-molecular weight heparin, or direct oral anticoagulants) ³⁴ <i>Patient deteriorates</i> without oxygenation + therapeutic anticoagulation, if alternative diagnoses are pursued (COPD exacerbation, heart failure exacerbation)	NT-proBNP (77) Albuterol-Ipratropium, Inhaled (77) Chest X-ray (63) Heparin IV (60)
Palpitations (R00.2) Unstable Paroxysmal Atrial Fibrillation with Rapid Ventricular Rate	66-year-old with a history of diastolic heart failure presenting with palpitations. Key Clinical Findings: tachycardia (rate >150 beats/min), hypotension, irregularly irregular pulse.	<i>Patient improves</i> with direct current cardioversion ³⁵ <i>Patient deteriorates</i> with IV nodal blockers (eg, metoprolol, diltiazem).	ECG 12-Lead (100) DCCV (100) Comprehensive Metabolic Panel (81) CBC with Differential (79) Consult to Cardiology (60) Troponin (56) Prothrombin Time/INR (100)
Hematemesis (K92.0) Acute Variceal Bleeding	59-year-old with a history of alcoholism and NSAID use presenting with hematemesis. Key Clinical Findings: tachycardia, mid epigastric pain, scleral icterus, spider angiomas.	<i>Patient improves</i> with fluid resuscitation, blood product administration, correction of coagulopathy with frozen plasma, proton-pump inhibitor, octreotide, and esophagogastroduodenoscopy ³⁶ <i>Patient deteriorates</i> without resuscitation, failure to correct coagulopathy, and esophagogastroduodenoscopy	Comprehensive Metabolic Panel (100) CBC with Differential (98) Consult to Gastroenterology (95) Type and Screen (95) Pantoprazole IV (91)

Last column reflects the most common clinical orders the test participants used in each case with the percent of occurrence in parentheses.

Abbreviations: CBC, complete blood count; COPD: chronic obstructive pulmonary disease; CSF, cerebrospinal fluid; DCCV, Direct Current Cardioversion; ECG, electrocardiogram; ICD, International Classification of Diseases; INR, international normalized ratio; IV, intravenous; NSAID, non-steroidal anti-inflammatory drug; R-CHOP, rituximab, cyclophosphamide, hydroxydaunorubicin, oncovin, prednisone.

Case-based scenarios

Table 1 summarizes key elements of the 5 simulated cases that participants were tested with.

Order search performance

The information retrieval performance of order search methods was assessed in terms of precision (positive predictive value: fraction of search result options that were ordered) and recall (sensitivity: fraction of orders that came from the search result options). When manually searching for orders by name or order sets, the user is

presented with $N(\text{manualOptions})$ to consider, of which we count a subset of $N(\text{uniqueManualOptions})$ after discounting repeats. Considering actual orders entered $N(\text{totalOrders})$ and $N(\text{manualOrders})$, Manual Search Precision is calculated as $N(\text{manualOrders})/N(\text{uniqueManualOptions})$ while Manual Search Recall is $N(\text{manualOrders})/N(\text{totalOrders})$. Respective metrics are calculated for the “search result” options and orders automatically presented by the recommender system. Note that when the recommender system was not available, $N(\text{manualOrders}) = N(\text{totalOrders})$ and thus Manual Search Recall = 100% since there was no alternative way to enter clinical orders. Further note that a user may be presented with a rec-

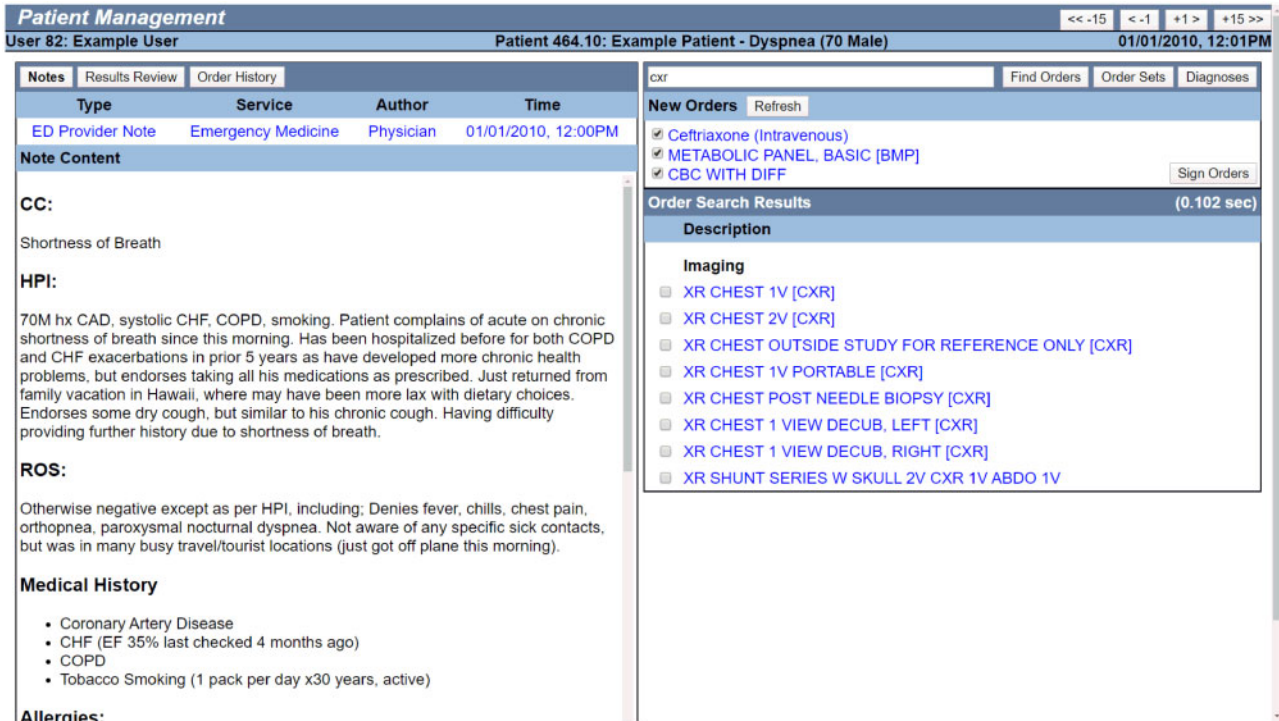


Figure 2. The simulated electronic medical record interface without the automated order recommender available. Note: Standard functions include navigation links to review notes and results (top left). Order entry includes a conventional search box for individual orders and pre-authored order sets (top-right). The New Orders selected but not yet Signed are presented in the top-right. The middle-right shows Order Search Results options after performing a manual search for individual orders based on the *cxr* prefix query entered in the search box above. Individuals could also use actual clinical order sets currently deployed at our institution with this interface (see Appendix for further examples).

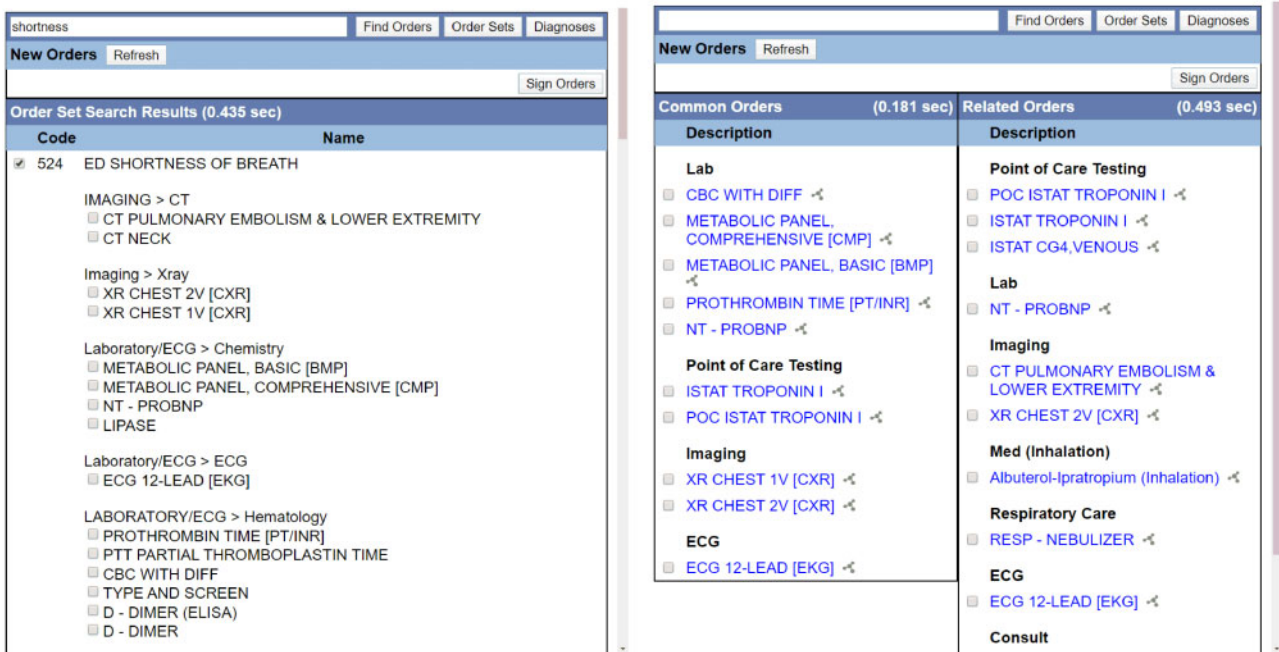


Figure 3. The simulated electronic medical record interface with order sets and automated order recommender available. Note: In all cases, participants had the option to manually search for pre-authored order sets or individual orders. The example on the left panel illustrates the user searching for *shortness* and then finding and opening the *ED Shortness of Breath* order set previously assembled by a hospital committee. For the physician-cases where the recommender option was turned on, rather than starting with a blank order search results field, the recommender algorithm dynamically presents a list of suggested clinical orders (right panel), in this example triggered by a presenting symptom code (Shortness of Breath, ICD9 786.05). Clinical orders predicted most likely to occur next are highlighted under *Common Orders* (top 10 when sorting options by positive predictive value relative to the available patient input data), while those under *Related Orders* are less likely but disproportionately associated with similar cases and thus may be more specifically relevant (top 10 options sorted by the negative log of the P-value association between the patient input data and suggested options). As users enter additional orders, the recommender algorithm continually updates the suggested lists based on the accumulating information.

Table 2. Primary and secondary outcomes

	Median (IQR) Recommender Off	Median (IQR) Recommender On	Additive effect (95% CI)
<i>Primary Outcome (Mean Score per Order)</i>	6.5 (5.3–7.6)	6.0 (5.1–7.5)	–0.11(–0.42–0.20)
	Median (Q1–Q3) Recommender Off	Median (Q1–Q3) Recommender On	Incidence Rate Ratio (95% CI)
<i>Ordering Outcomes</i>			
Total orders	15 (11–19)	16 (14–21)	1.09 (1.01–1.17)*
Total score	82 (64–101)	91 (72–112)	1.06 (1.01–1.12)*
Positively-scoring orders	11 (8.3–13)	12 (10–14)	1.08 (0.996–1.17)
Non-positively-scoring orders	2 (1–4)	3 (1–5)	0.99 (0.81–1.22)
<i>User-Experience Outcomes</i>			
Number of clicks	56 (36–72)	49 (35–65)	0.90 (0.83–0.99)*
Case completion time (min)	5.7 (4.0–8.6)	6.35 (4.3–8.6)	1.05 (0.94–1.19)

Medians and interquartile range (IQR) for each group are reported for summary context, but are not corrected for variation stemming from the simulated case or physician. Additive effect and incidence rate ratio estimates were calculated based on a linear mixed model with a random intercept for the clinician and a random intercept for the simulated case (see Statistical Methods section). Estimates for which the 95% CI does not overlap 1 are marked with a "*".

ommender option, ignore it, then subsequently manually order the same thing, which will be counted towards N(manualOrders) and not N(recommenderOrders).

Survey

After completion of the clinical scenarios, all participants were sent an electronic survey regarding their experience with the automated order recommender system used in this study. The survey used a 5-point Likert scale to assess physician opinions of this system and its potential impact on physician workflows (see Survey Responses below). Participants were also asked the following open-ended questions: “in which clinical scenarios would this system be most useful?” and “in which clinical scenarios would this system be least useful?” All open-ended responses were qualitatively coded and analyzed using thematic analysis.^{40,41}

Statistical Methods

Because several of the secondary outcome measures were nonsymmetrically distributed, median and interquartile ranges (IQR) are reported. The primary outcome (average score per order) was assessed using a linear mixed effects model with a normal distribution. Secondary outcomes were assessed using a generalized linear mixed effects model with a negative binomial distribution. All mixed effects models included a fixed effect for the recommender system (on/off), a random intercept for the physician, and a random intercept for the simulated clinical case (atrial fibrillation, gastrointestinal bleed, meningitis, neutropenic fever, and pulmonary embolism). No *P* values were reported for any mixed-effects model due to uncertainty in the statistical literature on the correct approach for calculating these values.⁴² In the secondary analyses, a result was considered nominally statistically significant if the confidence intervals for the incidence rate ratio estimate did not overlap a value of 1. For the secondary outcomes, negative and neutral scoring orders were pooled into a single “nonpositively scoring orders” analysis. This decision was made prior to analyzing the outcomes of the study. In the final order scores, there was substantial variation in the number of negatively scoring orders between the 5 simulated cases. In particular, there were no negatively scoring possible orders for the neutropenic fever case, and only 1 negatively scoring order for the pulmonary embolism case.

Therefore, negative orders were not individually modeled given the limited statistical power.

All statistical analyses were performed in R (Vienna, Austria). Linear mixed effects models were fit using the lme4 package.⁴³ The appropriateness of the negative binomial model was assessed with the visualizing categorical data package.⁴⁴ The ICC of the scores from individual expert reviewers was calculated with the incidence rate ratio package.⁴⁵

RESULTS

Participants

A total of 43 licensed physicians participated in this study, with a total of 215 unique observations. All participants received the intended treatment. The physicians had a median of 3.0 (IQR: [3.0–5.0]) years since obtaining their medical degree. The primary specialties of the participants were identified as Internal Medicine (including those with subspecialty training) for 32 (74%) participants, Emergency Medicine for 9 (21%) participants, and 1 each for Family Medicine and Pediatrics. Participants included 24 (56%) resident trainees and 19 (44%) attendings, who were board certified in their respective specialty.

Primary outcome

The median score per order for each physician-case was 6.2 (IQR: [5.2–7.5]). There was no significant difference detected in the mean score per order for simulated cases randomized to the automated order recommender (mean 0.11 decrease in score, 95% CI: [–0.41, 0.20]; Table 2). The random effects for physicians had an estimated standard deviation (SD) of 0.4 (95% CI: [0.1–0.6]), while the random effects for the simulated cases had an SD of 1.4 (95% CI: [0.7–2.7]). This suggests that more variation in the primary outcome may be attributable to the different simulated cases than to the specific physician participant.

Secondary outcomes

Ordering outcomes

The total scores per case were increased when the recommender system was available (median 82 vs 91 total points, incidence rate ratio 1.06, 95% CI: [1.01–1.12]; Table 2), but physicians placed more orders when the automated order recommender system was avail-

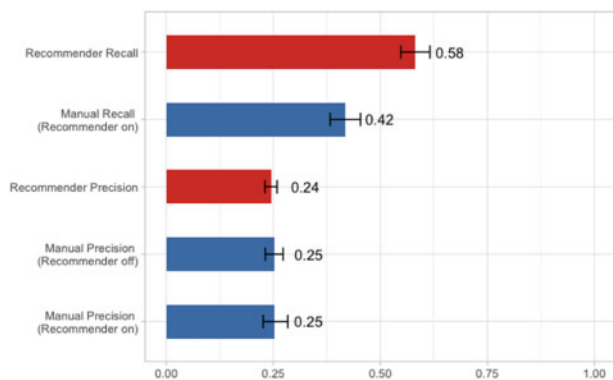


Figure 4. Order search performance metrics.

Note: Precision (positive predictive value: fraction of search results that were ordered) and recall (sensitivity: fraction of orders that came from the search results) for orders from the recommender system versus orders from the manual system. Error bars represent 95% bootstrap confidence intervals. Not adjusted for simulated case or physician variation. Note that recommender precision and recall are only defined with the recommender on, and manual recall is 100% by definition when the recommender is off because all orders must be made from manual search in that case.

Table 3. Physician survey responses

Survey Question	1	2	3	4	5
I would find the system useful in my job	0%	0%	5%	49%	47%
Using the system would make it easier to do my job	0%	5%	5%	44%	46%
This system would increase my productivity	0%	9%	5%	42%	44%
This system would let me complete tasks more quickly	0%	5%	7%	37%	51%
This system would increase my job performance	0%	7%	14%	47%	32%

Responses were assessed based on a 5-point Likert scale (1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree)

able (median 15 vs 16 orders per case; incidence rate ratio 1.09, 95% CI: [1.01–1.17]; Table 2). The number of clinically beneficial orders was not statistically significant between groups (median 11 vs 12 positive orders, incidence rate ratio 1.08, 95% CI: [0.996–1.17]). There was no difference in the number of clinically neutral or harmful orders for when the automated order recommender was available vs unavailable (median 2 vs 3, incidence ratio 0.99, 95% CI: [0.81–1.22]; Table 2).

User-experience outcomes

Overall, participants spent a median time of 6.0 minutes (IQR: [4.2–8.6]) per simulated case, with a median 51 total clicks (IQR: [36–69]). All physicians (100%) used the automated order recommender system at least once, including 127 (98%) of the 129 simulated cases where the order recommender was available.

Physicians made fewer total mouse clicks when the recommender was available (median 56 clicks without recommender vs 49 clicks with recommender; incidence rate ratio 0.90, 95% CI: [0.83–0.99]). Physicians did not take significantly more or less time to complete the cases with the automated order recommender system available (Table 2).

Average values for summary statistics (eg, case counts and primary and secondary outcomes) are included in Supplementary Tables, further stratified by clinical scenario and user status as resident vs nonresident and emergency medicine vs nonemergency medicine specialty.

Ancillary analyses

Order search performance

The precision and recall of the automated order recommender vs manual search are shown in Figure 4. The precision of the manual search vs the recommender system was comparable, suggesting that users had to sift through roughly equivalent numbers of irrelevant options to find the clinical orders they wanted. The recall of the recommender system was greater than manual search, indicating users were more likely to find the clinical orders they wanted from the automated recommender lists than from options returned by manual search.

Survey responses

All 43 participants responded to the survey (100%). Overall, the clinical decision tool was positively received by the study participants, where 96% agreed or strongly agreed that the tool would be useful for their position (Table 3). Moreover, 90% agreed or strongly agreed that the system would make their job easier, and 86% felt that it would increase their productivity. Thematic analysis of the open-ended questions revealed that the majority of respondents felt the system would be useful for patients who have a clear diagnosis or whose clinical problems could be guided by a stepwise, algorithmic approach. Others mentioned the tool's utility for diseases that may require several simultaneous orders (eg, diabetic ketoacidosis). Additional comments indicate physicians felt that the tool would be less useful for subspecialized care or for patients that require few simultaneous orders.

DISCUSSION

An automated order recommender system interface for common clinical scenarios seen in hospital medicine and emergency medicine was usable and accepted by physicians, without adversely affecting the quality of patient care decisions reflected in adverse or irrelevant orders. The automated order recommender system reduced the number of clicks per case and did not affect the amount of time physicians spent on the simulated EHR interface. Physicians placed more orders overall (median 1.0 additional order per simulation). The automated order recommender demonstrated superior recall of orders, suggesting that users were more likely to find the orders they wanted from the automated order recommender rather than from manual searches. The tool was positively received by the study participants, who identified clear benefits toward their workflow and productivity. This represents a key study to examine the usability and acceptability of clinical recommender decision support tools on physician ordering habits and patient care.

There is wide variability in clinical practice, even in instances with clear guideline-directed diagnostic and treatment algorithms for well-defined clinical problems.¹ Such variability may compromise care quality, cost effectiveness, or expedient healthcare delivery.² Healthcare systems have sought to improve both the quality of patient care and the EHR experience by providing standardized order sets, which can promote evidence-based care.^{46–49} Healthcare systems have also sought to use EHR alerts to encourage evidence-

based ordering of diagnostic tests in both inpatient and outpatient settings.^{50,51} Given the variability of clinical practice,¹ guidance for up-to-date medical care must come from multiple sources. Machine learning tools have the potential to augment these existing features, and they offer a potential advantage of a collaborative filtering approach that can rapidly and automatically adapt to newly emerging practices. Our automated order recommender essentially functions as a dynamic order set that continuously updates in response to new patient information, demonstrating increased accuracy and reduced need for conventional manual searches. While there are challenges to designing and maintaining such a system,⁵² there may be several benefits, including increased physician acceptance and usage. In this study, physicians could use any of our institution's order sets for the simulated cases, regardless if the recommender was available or not. Notably, 100% of physicians in this study used the recommender options at least once, even though it was completely optional. The automated order recommender system interface received largely positive views by our participants, suggesting that physicians will accept machine-generated clinical order tools if they are embedded into clinical workflows. Further studies are needed to evaluate the role of machine-learning tools in EHR interfaces and how they can augment existing features that promote evidence-based care.

There are several potential consequences to an automated recommender system. First, it may lead physicians astray by following "common, but not necessarily good" practices. Our expert panel found no significant deterioration in the overall quality of clinical orders, but there was still substantial case-variability in the amount of clinically appropriate orders placed with or without the automated recommender system. These findings highlight the ongoing variability of clinical practice among physicians (even when additional point-of-care tools are given to them). A second concern of automated order recommender systems is they may increase extraneous cognitive load by bombarding the user with continuously updated order suggestions (some of which may not be relevant). Indeed, time-motion studies indicate that clinicians already spend most of their time in the EHR,^{53,54} with many spending significant time searching for and entering orders.⁵⁵ Our study showed a reduction on reliance of manual searches and navigational clicks with an automated recommender system. However, there was no reduction in the amount of time that physicians spent per simulated case. The simulated test setting with a predetermined duration and number of cases may have led participants to artificially "fill" the time within cases, or perhaps the reduction in manual search efforts freed their cognitive attention to attend more to the medical decision-making tasks of each case. Additionally, other authors have shown that most of a clinician's time in the EHR is spent in data review (reviewing clinical notes, laboratory results, or diagnostic reports),⁵⁵⁻⁵⁷ which was simplified in these clinical scenarios. A third limitation regarding automated order recommender systems is that they can present users with a continuously updated menu of order options, which may result in excessive ordering and increased healthcare expenditures. Our study found that users placed an additional 1.0 order per case with the automated recommender, which could result in significant increases in healthcare costs when viewed on a macro level. In contrast, order sets have been shown to promote cost-effectiveness.^{46,47} Future studies should consider the implementation of automated order recommender systems in real practice environments to assess whether they result in time savings for physicians navigating the EHR without compromising the cost or quality of care.

There are several limitations to this study. Our tool was based on a clinical data warehouse of EHR data that may not be available at all institutions. This study used simulated cases, and it remains to

be seen how an automated order recommender system would perform in "live" practice settings with real patients. The simulated cases pertained to inpatient or emergency medicine contexts, which limits the generalizability of these findings. Similarly, our primary users were internal medicine physicians, and it is unclear how physicians from other specialties would respond to this system. The recommender system was based on data collected from 2009-2014, and as such, clinical practice patterns may have changed. Our users were given an orientation of the recommender system and its purpose before engaging with the practice scenarios, which likely contributes a Hawthorne effect on how users interacted and viewed the system.⁵⁸ Each testing session was prescheduled for a fixed time (1 hour for orientation, 5 test cases, and survey), which may have artificially constrained the variability in task completion time. Although our expert panel used previously validated methodology to devise a scoring system,³⁹ the initial interrater agreement was moderate for some clinical orders, which limits the generalizability of these findings to other clinical scenarios or healthcare settings. Finally, the lack of a broadly accepted, open-architecture platform that allows for custom workflow integrations into common commercial EHRs limits the ability to easily implement systems similar to the 1 used in this study.

At a time when the EHR is met with distrust and negativity by clinicians due to the burdens of documentation and data entry, automated order recommender systems represent a key opportunity to improve the quality, consistency, and experience of healthcare. This study represents an important step towards a future where EHRs anticipate clinical needs without users even having to ask, so that clinicians can start to feel like the computers are working for them, instead of the other way around.

CONCLUSIONS

Clinical order suggestions from a data-driven recommender system were readily used and accepted by physicians across a variety of simulated clinical cases. The clinical appropriateness of orders entered were comparable when supported by automated recommendations, even as the system increased the number of clinical orders placed per case. Physicians were more likely to find the clinical orders they wanted using such tools as compared to manual search methods (ie, superior recall), and reduced the number of mouse clicks, but did not change the overall amount of time they spent in a simulated EHR setting. Clinicians overall viewed such clinical recommender systems positively, perceiving a clear potential benefit toward their workflow.

FUNDING

This research was supported in part by the NIH Big Data 2 Knowledge initiative via the National Institute of Environmental Health Sciences under Award Number K01ES026837, the Gordon and Betty Moore Foundation through Grant GBMF8040, and a Stanford Human-Centered Artificial Intelligence Seed Grant. This research used data or services provided by STARR, Stanford medicine Research data Repository, a clinical data warehouse containing live Epic data from Stanford Health Care, the University Healthcare Alliance and Packard Children's Health Alliance clinics and other auxiliary data from Hospital applications such as radiology PACS. The STARR platform is developed and operated by Stanford Medicine Research IT team and is made possible by Stanford School of Medicine Research Office. The 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

Drafting or critically revising manuscript: AK, RA, JHC. Statistical analysis: RA, MB. Design and execution of user testing protocol: JC, DM, DS, MM. Authoring and expert evaluation of clinical content: AK, JH, LS. Conception or design of the overall study: RA, MKG, SA, JHC.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

JHC is co-founder of Reaction Explorer LLC that develops and licenses organic chemistry education software and has been paid consulting or speaker fees from the National Institute of Drug Abuse Clinical Trials Network, Tuolac Inc., Roche Inc., and Younker Hyde MacFarlane PLLC. All other authors have no competing interests to declare.

REFERENCES

- Richardson WC, Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: National Academy Press; 2001.
- Tricoci P, Allen JM, Kramer JM, Califf RM, Smith SC Jr. Scientific evidence underlying the ACC/AHA clinical practice guidelines. *JAMA* 2009; 301 (8): 831–41.
- Health and Human Services Department. 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. *Fed Regist* 2012; 77(177): 54163–292. <https://www.federalregister.gov/d/2012-20982> Accessed May 6, 2020.
- 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 (9): 759–60.
- Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. *N Engl J Med* 2011; 365 (19): 1758–9.
- Longhurst CA, Harrington RA, Shah NH. A “green button” for using aggregate patient data at the point of care. *Health Aff* 2014; 33 (7): 1229–35.
- Committee on the Learning Health Care System in America, Institute of Medicine. Smith M, Saunders R, Stuckhardt L, Michael McGinnis J (eds.). *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America*. Washington, DC: National Academies Press; 2013.
- Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Aff* 2014; 33 (7): 1163–70.
- Holroyd BR, Bullard MJ, Graham TAD, Rowe BH. Decision support technology in knowledge translation. *Acad Emerg Med* 2007; 14 (11): 942–8.
- 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 (12): 1409–16.
- Overhage JM, Tierney WM, Zhou XH, McDonald CJ. A randomized trial of “corollary orders” to prevent errors of omission. *J Am Med Inform Assoc* 1997; 4 (5): 364–75.
- Bates DW, Kuperman GJ, Wang S, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *J Am Med Inform Assoc* 2003; 10 (6): 523–30.
- Li RC, Wang JK, Sharp C, Chen JH. When order sets do not align with clinician workflow: assessing practice patterns in the electronic health record. *BMJ Qual Saf* 2019; 28: 987–96. <http://dx.doi.org/10.1136/bmjqs-2018-008968>.
- Kumar A, Allaudeen N. To cure sometimes, to relieve often, to comfort always. *JAMA Intern Med* 2016; 176 (6): 731–2.
- Sittig DF, Wright A, Osheroff JA, et al. Grand challenges in clinical decision support. *J Biomed Inform* 2008; 41 (2): 387–92.
- Middleton B, Sittig DF, Wright A. Clinical decision support: a 25 year retrospective and a 25 year vision. *Yearb Med Inform* 2016; 25 (S 01): S103–16.
- 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 Inform Assoc* 2017; 24 (3): 472–80.
- Zhang Y, Levin JE, Padman R. Data-driven order set generation and evaluation in the pediatric environment. *AMIA Annu Symp Proc* 2012; 2012: 1469–78.
- Klann J, Schadow G, McCoy JM. A recommendation algorithm for automating corollary order generation. *AMIA Annu Symp Proc* 2009; 2009: 333–7.
- Wright AP, Wright AT, McCoy AB, Sittig DF. The use of sequential pattern mining to predict next prescribed medications. *J Biomed Inform* 2015; 53: 73–80.
- Chen JH, Goldstein MK, Asch SM, Altman RB. *Usability of an Automated Recommender System for Clinical Order Entry*. AMIA; 2016. <https://knowledge.ama.org/ama-63300-1.3360278/t001-1.3365273/f001-1.3365274/2497982-1.3365629/2496685-1.3365624?qr=1> Accessed May 6, 2020.
- Chen JH, Podchiyska T, Altman RB. OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records. *J Am Med Inform Assoc* 2016; 23 (2): 339–48.
- King AJ, Cooper GF, Hochheiser H, Clermont G, Hauskrecht M, Visweswaran S. Using machine learning to predict the information seeking behavior of clinicians using an electronic medical record system. *AMIA Annu Symp Proc* 2018; 2018: 673–82.
- Hunter-Zinck HS, Peck JS, Strout TD, Gaehde SA. Predicting emergency department orders with multilabel machine learning techniques and simulating effects on length of stay. *J Am Med Inform Assoc* 2019; 26 (12): 1427–36.
- Emami S, Ting DY, Healey M, Lipsitz SR, Karson AS, Bates DW. Physician beliefs about the meaningful use of the electronic health record: a follow-up study. *Appl Clin Inform* 2017; 08 (04): 1044–53.
- Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. *JAMA* 2018; 319 (1): 19–20.
- Gawande A. Why doctors hate their computers. *The New Yorker* [Internet]. 2018 Nov 5 [cited 2019 Aug 15]; <https://www.newyorker.com/magazine/2018/11/12/why-doctors-hate-their-computers> Accessed May 6, 2020.
- 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* 2015; 2015: 435–9.
- Wang JK, Hom J, Balasubramanian S, et al. An evaluation of clinical order patterns machine-learned from clinician cohorts stratified by patient mortality outcomes. *J Biomed Inform* 2018; 86: 109–19.
- Linden G, Smith B, York J. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Comput* 2003; 7 (1): 76–80.
- Smith B, Linden G. Two decades of recommender systems at Amazon.com. *IEEE Internet Comput* 2017; 21 (3): 12–8.
- Freifeld AG, Bow EJ, Sepkowitz KA, et al. Clinical practice guideline for the use of antimicrobial agents in neutropenic patients with cancer: 2010 update by the Infectious Diseases Society of America. *Clin Infect Dis* 2011; 52 (4): e56–93–e93.
- Tunkel AR, Hartman BJ, Kaplan SL, et al. Practice guidelines for the management of bacterial meningitis. *Clin Infect Dis* 2004; 39 (9): 1267–84.
- Konstantinides SV, Barco S, Lankeit M, Meyer G. Management of pulmonary embolism: an update. *J Am Coll Cardiol* 2016; 67 (8): 976–90.
- January CT, Wann LS, Alpert JS, et al. 2014 AHA/ACC/HRS Guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force

- on Practice Guidelines and the Heart Rhythm Society. *J Am Coll Cardiol* 2014; 64 (21): e1–76.
36. Garcia-Tsao G, Sanyal AJ, Grace ND, Carey W, Practice Guidelines Committee of the American Association for the Study of Liver Diseases, the Practice Parameters Committee of the American College of Gastroenterology. Prevention and management of gastroesophageal varices and variceal hemorrhage in cirrhosis. *Hepatology* 2007; 46 (3): 922–38.
 37. Chiang J, Kumar A, Morales D, et al. Physician usage and acceptance of a machine learning recommender system for simulated clinical order entry. *AMIA Summits Transl Sci Proc* 2020; 2020: 89–97.
 38. 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.
 39. Hsu C-C, Sandford BA. The Delphi technique: making sense of consensus. *Pract Assess Res Eval* 2007; 12 (10): 1–8.
 40. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol* 2006; 3 (2): 77–101.
 41. Chun Tie Y, Birks M, Francis K. Grounded theory research: A design framework for novice researchers. *SAGE Open Med* 2019; 7: 205031211882292.
 42. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. *Am Stat* 2016; 70 (2): 129–33.
 43. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models using lme4 [Internet]. *arXiv [Stat.CO]* 2014. <http://arxiv.org/abs/1406.5823> Accessed May 6, 2020.
 44. Zeileis MD. vcd: visualizing categorical data. *R Package Version* 2020; 1: 4–7 Accessed May 6, 2020.
 45. Gamer M, Lemon J, Fellows I, Singh P. Coefficients of Interrater Reliability and Agreement for quantitative, ordinal and nominal data. *R Package Version* [Internet]. 2019 [cited 2020 Jun 1]; 0.84.1. <https://cran.r-project.org/web/packages/irr/index.html> Accessed May 6, 2020.
 46. Brown KE, Johnson KJ, DeRonne BM, Parenti CM, Rice KL. Order set to improve the care of patients hospitalized for an exacerbation of chronic obstructive pulmonary disease. *Ann ATS* 2016; 13 (6): 811–5.
 47. Radosevich MA, Wanta BT, Meyer TJ, et al. Implementation of a goal-directed mechanical ventilation order set driven by respiratory therapists improves compliance with best practices for mechanical ventilation. *J Intensive Care Med* 2019; 34 (7): 550–6.
 48. Nichols KR, Petschke AL, Webber EC, Knoderer CA. Comparison of antibiotic dosing before and after implementation of an electronic order set. *Appl Clin Inform* 2019; 10 (02): 229–36.
 49. Zeidan AM, Streiff MB, Lau BD, et al. Impact of a venous thromboembolism prophylaxis “smart order set”: improved compliance, fewer events. *Am J Hematol* 2013; 88 (7): 545–9.
 50. Chin K-K, Hom J, Tan M, et al. Effect of electronic clinical decision support on 25(OH) vitamin D testing. *J Gen Intern Med* 2019; 34 (9): 1697–9.
 51. Jun T, Kwang H, Mou E, et al. An electronic best practice alert based on choosing wisely guidelines reduces thrombophilia testing in the outpatient setting. *J Gen Intern Med* 2019; 34 (1): 29–30.
 52. Chen JH, Alagappan M, Goldstein MK, Asch SM, Altman RB. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. *Int J Med Inform* 2017; 102: 71–9.
 53. Desai SV, Asch DA, Bellini LM, et al. Education outcomes in a duty-hour flexibility trial in internal medicine. *N Engl J Med* 2018; 378 (16): 1494–508.
 54. Kumar A, Chi J. Duty-hour flexibility trial in internal medicine. *N Engl J Med* 2018; 379 (3): 300.
 55. Ouyang D, Chen JH, Hom J, Chi J. Internal medicine resident computer usage: an electronic audit of an inpatient service. *JAMA Intern Med* 2016; 176 (2): 252–4.
 56. Chi J, Bentley J, Kugler J, Chen JH. How are medical students using the Electronic Health Record (EHR)? An analysis of EHR use on an inpatient medicine rotation. *PLoS One* 2019; 14 (8): e0221300.
 57. Wang JK, Ouyang D, Hom J, Chi J, Chen JH. Characterizing electronic health record usage patterns of inpatient medicine residents using event log data. *PLoS One* 2019; 14 (2): e0205379.
 58. McCambridge J, Witton J, Elbourne DR. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *J Clin Epidemiol* 2014; 67 (3): 267–77.