A Data Driven System for Clinical Preventive Order Recommendations

Linda Zhang, BS[1], Colin Walsh, MD, MA[1], Daniel Fabbri PhD[1]
[1]Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee

Abstract
Optimal use of preventive care could prevent over 50,000 deaths per year in those under 80 years of age. Although guidelines have been established, rates of use of preventive care are below recommended levels. Increased use of these services can be induced by reminding physicians with alerts and clinical decision support. Initially defining and updating the content for the alerts and finding where they fit in ordering workflow is time consuming, deterring their use. We attempt to solve this problem by creating a data-driven system that automatically learns current preventive care ordering practices. We designed a framework for preventive order recommendation using machine learning. We refined this system and a study cohort using colonoscopy order histories. The colonoscopy classifier output by the system achieves a receiver operating characteristic area under the curve of 0.893. In comparison to existing guidelines, the classifier predicts orders at a similar sensitivity but higher specificity. These results are encouraging, indicating that this framework for learning preventive order recommendations is feasible.

Introduction
Preventive medicine focuses on preventing disease rather than treating it[1]. It achieves this by using practices such as health promotion, immunization, and screening[1]. Currently, many preventative orders exist in medical systems. Some examples include colonoscopies and mammograms, which are used for colon and breast cancer screening. The guidelines that specify when to place preventive orders for screening purposes can be complex, often involving a systematic review of clinical evidence. These guidelines are created by experts and updated over time[2]. Although guidelines have been established, rates of preventive service use are below recommended levels[3]. One prominent example of this can be seen in colorectal cancer screening, for which only 59.1% of adults had been screened in 2010[4]. Farley et al. estimated in 2010 that optimal use of nine of the most common preventive measures could prevent an additional 50,000-100,000 deaths per year in those under 80 years of age[5].

Increased use of preventive services can be induced by reminding physicians when to give preventive care using alerts and clinical decision support in the computerized order entry (CPOE) system, leading to improved practitioner performance[6]. There are, however, challenges and associated costs with integrating decision support and guidelines into clinical practice via CPOE[7]. Two major challenges are 1) the manual maintenance of decision support content and guidelines, and 2) the practicality of the decision support at the time of use. The first challenge pertains to keeping the guidelines up to date and relevant. Updates to guidelines, which decision support requires as input, necessitate an administrator to periodically check and modify each guideline instance. The second challenge involves the practicality of the clinical decision support timing. In situations such as acute care, it may not be appropriate to remind the provider to schedule a preventive screening. As a result of these challenges, clinical decision support has not always been implemented when it can and should be[8].

One popular method used to generate recommendations from data is machine learning. Machine learning has been used in many fields to discover patterns in data sets[9]. Since electronic medical record (EMR) data reflects physician perceived guidelines used to place orders, research has utilized the data to automatically generate decision support content[10,11]. Studies have used techniques such as Bayesian networks to analyze local order-entry data[10], frequent item set and association rule mining to learn order sets and corollary orders [12], and market-basket analysis with natural language processing to provide order recommendations[13]. These methods all focus on producing order sets and corollary orders. Our study focuses on preventive orders as opposed to order sets, and uses a binary classification method instead of trying to group orders. Similar to the other studies, we do not attempt to learn new information or guidelines, but seek to learn when orders are placed, and build a system that will predict them.

In this paper, we attempt to solve the problem of automatically recommending preventative clinical orders to physicians. We use supervised machine learning to develop a system which mines clinical ordering patterns (specifically targeting preventive orders), automatically learning existing practices for when orders should be placed based on various clinical and operational features using data in the EMR. The system analyzes current ordering patterns made by physicians, using the operational features to predict the practicality of ordering at each appointment. Specifically, we analyze non-traditional operational meta-data such as appointment length, appointment location, and days between appointment and scheduled date in the classification model. This system is

generalizable in that it is able to learn the ordering practices specific to a given medical center, and update periodically as guidelines change. The trained system can be used to recommend preventive orders at future appointments for patients that satisfy the learned criteria.

Methods

Overview

In this study, we describe a system that uses supervised machine learning to build a model to predict whether a preventive order should be placed in a patient's upcoming appointment. We treat the prediction as a supervised classification problem, and classify orders in a binary fashion. Features for the classifier were chosen from observations that were made regarding clinical guidelines and patterns in workflow for preventive orders. The system works as follows: 1) a target order is specified, 2) clinical and operational features along with order history are input to the classifier, 3) the classifier is trained on the data set, and 4) the classifier's predictions are evaluated on retrospective order histories. Using random forests generated from the input data, the output classifier can then predict if the order for specific patients is placed (i.e., the recommendation). The target order that we use to test our system is colonoscopy, as a preventive order for colorectal cancer screening (Fig. 1).

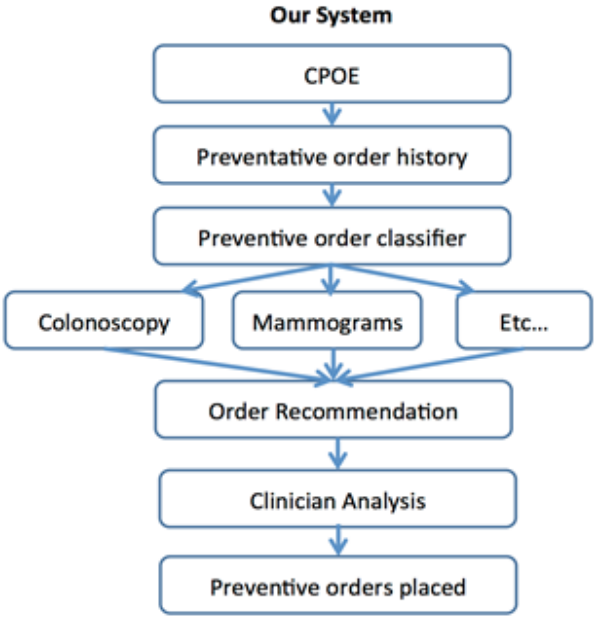

Figure 1. Overview of our target system architecture; our test order is colonoscopy.

Setting

This study was conducted using data from the Vanderbilt University Medical Center (VUMC) in middle Tennessee. The subjects of the data used in this study are patients in 2013 at the VUMC for whom physicians placed outpatient orders. This study was approved by Vanderbilt's IRB.

Data

The data used as input to the classifier consisted of patient data such as demographics, medical history and appointment data. The data came from the Vanderbilt Outpatient Order Management (VOOM) system and StarPanel system. VOOM is Vanderbilt's self-developed order entry system for outpatient orders. StarPanel is Vanderbilt's homegrown electronic medical record system[14]. Since we focused on predicting the outcome of ordering during specific appointments, we separated data into "events", or appointments that consisted of information about the patient, their history, and logistics about the appointment taken from both sources. Data from VOOM provides the order-entry data, patient demographics and the orders' International Classification of Diseases (ICD9) codes. StarPanel additionally provides appointment information, clinical notes, ICD9 code history and current procedural terminology (CPT) code history. The StarPanel data was joined with the medical record numbers and dates within 5 days from the VOOM data set. A constraint on date ensured that the order was made after the appointment date.

The input for the system included patient data and clinical operations data. Patient data, such as demographics and history, provide features that generate decisions close to those of established guidelines. Operational information (e.g. appointment length, location, and length of time from the date the appointment was made) is important in providing insight into where the order fits in the ordering workflow. It is not always practical to order preventive orders in every appointment. If screening alerts were generated for every appointment, then these alerts may hinder instead of help the clinicians find the orders that are relevant. However, because our goal is to predict which patients were likely candidates for the specified order, we avoided using information from the day of the appointment, such as patient complaints and other orders made that day. This information is unused because it correlates with the patient's diagnosis for that day, not with flags for preventive care.

Feature analysis and selection

We selected classes of features for the classifier that we believed would capture the guidelines for preventive orders. All features could only include information that is available before the appointment in which the order was placed. Since our test order was colonoscopy, we started with the feature classes of age, time from last procedure, and

clinical note keywords (e.g., colorectal cancer, polyps) from colonoscopy guidelines[15]. We then added operational features that we believed would give insight into the appropriate timing of the order in the workflow. All feature classes were tested for how well they helped to predict the order. We used the following feature classes:
Age: The age of the patient in years. We chose age because preventive orders are often scheduled based on age.
Time from last procedure: The last time the procedure/test associated with the order was performed in years, maxed at 15. We chose this feature because guidelines recommend colonoscopies once every 5-10 years [15].
ICD9 parent code: All ICD9 parent codes in the medical history of the patient. Each ICD9 code is a binary feature.
Clinical note: The last time since a keyword was seen in clinical notes in the medical history of the patient, measured in years, maxed at 15. Each keyword is one feature, and keywords must be manually specified. We selected keywords that were indicators of high risk for colorectal cancer (Appendix A) [15].
Days from date scheduled: The days between the date the appointment was scheduled and date of appointment. We chose to test this feature because we observed that appointments made for sudden problems (e.g., injury, illness) were not likely to have preventive orders. We hypothesized that appointments created over a month in advance are more likely to be checkups, in which an order made for the purpose of screening has a better chance of being placed.
Appointment length: The length of time in minutes that the appointment is scheduled for. We chose to test this feature because we observed that the pre-assigned lengths of time for check-up appointments differ from those of appointments for specific problems. We hypothesized that preventive orders were more likely to be given at check-up appointments, which would have a different amount of time pre-allocated.
Appointment location: The location of appointment from 88 possible locations. Each location is a binary feature.

The input of a classification system is a feature matrix and a target or outcomes matrix. Each row in the feature matrix refers to an event, denoted by a medical record number and date. Each column represents a different feature. The features we used were binary (e.g., locations: was at location or not) and real-valued (e.g., age). The corresponding row in the outcomes matrix is a binary representation of whether the order was made that day or not. By this design, a patient can have more than one entry in the matrices if they had multiple appointments.

We created a classifier to test each of the listed classes of features, and compared their area under the receiver operating characteristic (ROC) curve (AUC) scores. Next, we combined the features into one combined classifier, and removed or kept the features that we deemed to be important to the classifier by examining the AUC of different combinations of the classes of features and comparing Gini importance values (e.g., weighted mean value describing classification accuracy as random forest nodes are split)[16].

Classification
Python's scikit-learn package for machine learning was used to develop the classifiers for orders. Many different methods can be used for classification, including Naive Bayes, logistic regression and random forests. We chose to use random forests as a result of preliminary experiments, in which we found that the relationships between the data are inherently nonlinear. The random forest used 100 estimators, a minimum sample size of 20 per node, and a maximum depth of 5. Since random forests never compare features to one another, normalization is unnecessary. To tune our system, we chose to use colonoscopy as our test order. We analyzed the performance of the 7 feature classes in individual classifiers and a combined classifier, and removed features that did not improve prediction.

Input refinement
Initially, we tested the system on the entire population. However, after a preliminary analysis, we discovered the cohort required refinements in large part because orders were not always placed in a preventive setting. Therefore, we filtered the population to target events in which patients had a colonoscopy as a preventive order by removing events from the gastrointestinal (GI) medicine and gastroenterology departments (Table 1). Colonoscopy orders placed from these departments were likely to be for non-preventive reasons. After additional analysis, we observed that some locations never placed certain preventive orders such as colonoscopy, potentially inflating the true negative prediction rates. We therefore developed a second subpopulation in which we removed events from locations that had never ordered colonoscopies previously, resulting in 14 remaining locations (Table 1). To analyze these refinements, we compared

Table 1. Population and subpopulations that were used as input for the system.

Population	Number of Events
Entire VOOM population	106869
Population where appointments with orders from the departments GI medicine and gastroenterology are removed	80782
Population restricted to locations from which colonoscopy is ordered	50258

the resulting AUC of the final classifier on the subpopulations and total population. Specifically for the individual location classifier, we examined the sensitivity and specificity of the predictions, oversampling the positive class events to be equal to the negative (increasing prior probability) and using a threshold of $p \geq 0.5$.

Classifier evaluation
We evaluated the performance of the colonoscopy classifier by using 5-fold cross validation and calculating the average ROC AUC scores. In our cross validation, the folds were created by shuffling rows, keeping rows from the same patient together. By doing this, we ensured that all events by one patient were placed in the same folds to keep the results consistent for each patient.

Comparison to existing guidelines
We compared the performance of our classifier to the performance of existing base guidelines for colonoscopy screening by measuring the sensitivity and specificity of predicting ordered colonoscopies. For these predictions, we oversampled the positive class events to equal the negative class and used a threshold of $p \geq 0.5$. The base guidelines we used as a model for comparison included a check for the age of the patient within a range of 50-75 and the number of years since the last colonoscopy was performed (10≤ if no history in the clinical note included keywords in Appendix A, 5≤ if a history was found).

Results
Characteristics of Data Set
The data from VOOM was taken in the 2013 year. During this time period 585,000 orders were placed by 1,000 clinicians, representing 63,027 distinct patients. The VOOM data was joined with data from StarPanel to obtain additional information. The summary statistics for this joint data set are presented (Table 2).

Table 2. Joint data summary statistics

Total number of distinct patients	58550	Total number of colonoscopy orders	1454
Total number of distinct orders	1335	Average number of orders per appointment	5.733
Total number of appointments	87616	Average number of colonoscopies per appointment	0.017

Classification results
Table 3 shows the results for each feature class used to construct its own classifier, tested on the total population. The results show the location classifier performed the best followed by the age and ICD9 classifiers.

Table 4 details the classification results on various populations. The combined classifier using the original 7 classes of features on the total population had an AUC of 0.819 (Table 4). After removing the gastro-related departments, the AUC of the combined classifier increased to 0.865 (Table 4). In addition, the AUCs of the individual classifiers were constant or increased (Table 3).

Table 3. Mean AUC for each separate classifier for colonoscopy using 5-fold cross validation

Classifier	**Number of Features**	**Total Population AUC**	**Removed Department Events AUC**
Age	1	0.680	0.774
Time from last procedure	1	0.540	0.563
ICD9 parent code history	649	0.684	0.771
Clinical note	12	0.590	0.627
Days from date scheduled	1	0.587	0.591
Appointment length	1	0.647	0.708
Appointment location	88	0.850	0.858

Table 4. Mean AUC for combined classifier for colonoscopy using 5-fold cross validation.

Classifier	**AUC**
Classifier with 7 feature classes, on total population	0.819
Classifier with 7 feature classes, on subpopulation where events from gastro-related departments removed	0.865
Classifier with 6 feature classes (ICD9 parent code history features removed), on subpopulation where events from gastro-related departments removed	0.923
Classifier with 6 feature classes (ICD9 parent code history features removed), on subpopulation where events from gastro-related departments and locations where colonoscopy were never ordered removed	0.893

We examined the features with the highest impact on the classifier (by sorting the features by the Gini importance value). The top ten features sorted by their impact on the prediction using Gini importance are presented (Table 5). Age and location were found to have the highest impact on prediction. Additionally, we observed that the ICD9 parent code history class contributed 649 features, all of which had very low importance separately. After removing the ICD9 features, the AUC increased to 0.923 (Table 4).

Table 5. Top 10 features sorted by Gini importance.

Feature	**Gini importance**
Age	0.225
Location: VECL (Internal Medicine)	0.128
Appointment length	0.104
Days from date scheduled	0.074
Location: OAKS (Internal Medicine)	0.065
Location: J071 (Internal Medicine)	0.064
Location: A044 (Internal Medicine)	0.055
Time from last procedure	0.042
Location: A015 (General Internal Medicine)	0.036
Location: A041 (Internal Medicine)	0.036

The six classes of features in the final classifier were age, time from last procedure, days from date scheduled, clinical note, appointment length and appointment location. The high AUC of the appointment location classifier led to additional testing for location features. Therefore, we created a subpopulation in which all events at locations from which colonoscopies were never ordered were removed. The resulting population included 14 locations. While the combined classifier run on this subpopulation's AUC decreased to **0.893** (Table 4) and the individual location class classifier's AUC decreased to 0.760, the AUC was still highly predictive. By examining the events, we found that 6 locations ordered 85% of all the colonoscopies, inflating the predictive value of location. Using an oversampled positive class and $p \geq 0.5$ threshold, we found that the location classifier achieves a sensitivity of 0.860 and specificity of 0.540. The classifier predicts positive for the 6 locations before mentioned, and achieves a high sensitivity at the cost of a low specificity.

Comparison to existing guidelines

We found that a model based on existing guidelines produced predictions with a sensitivity of 0.890 and a specificity of 0.239. Our model in comparison produced a sensitivity of 0.893 and a specificity of 0.750.

Discussion

In this study, we designed a system that creates a classifier for a given preventive order. Through testing, we found that a classifier using the feature classes age, last procedure, days from date scheduled, clinical note, appointment length and appointment location produced the best AUC. We compared the predictions from our classifier to those of a model based on existing guidelines, and found that it outperformed the model with a higher specificity.

To target colonoscopies ordered for preventive screening, we created a subpopulation in which we 1) removed events where orders were made from the GI medicine and gastroenterology departments (these orders were often not made for preventive reasons), 2) removed the ICD9 parent code history features, because we found that those features actually lowered the combined classifier's AUC, and 3) removed clinic locations that never ordered colonoscopies (without removing these extra locations, we would unfairly attain a high true negative rate and a high AUC even though the specificity was extremely low). Through the comparisons of the Gini importance of features, we found that age and location were very prominent features.

The combined classifier performed the best in the subpopulation with events from gastro-related departments removed. In the smaller subpopulation in which events from locations where colonoscopy is never ordered are removed, the classifier performs worse. This occurs because there are a disproportionate number of locations from which colonoscopies are never ordered, allowing a few locations to have a high impact on the results. Removing the locations in which colonoscopies are never ordered removes a significant number of true negative classifications. We conclude that location is an important operational feature for Vanderbilt, though it may not be so in other medical centers. Overall, we found that the classifier detected clinical features useful for predicting colonoscopy orders: age, time from last procedure, and clinical note keywords. We were able to target appointments in which preventive colonoscopy orders are more likely to be ordered by including the operational features appointment location, appointment length, and days since date scheduled.

From examining the misclassifications of our model, we found that a majority of the false positives are cases in which the patient does fit the mandated guidelines, and fit the operational guidelines that our model discovered. These may be cases in which preventive screening should have been ordered, but was not. Of the false negatives in

our final model, we found that the greatest number of the false negative colonoscopies resulted from diagnostic colonoscopies performed for reasons such as blood in stool, but were not ordered by gastro-related departments.

Comparing to a model created from existing guidelines for colonoscopy, we found that sensitivity of our model was equal to that of existing guidelines, but our model had a much higher specificity. This indicated that our model may be able to predict the appropriate times to recommend a colonoscopy based on the included operational features, which reduces the false negative misclassifications greatly. This can translate to CDS alerts which are timely placed.

There are several limitations of this framework, however. One limitation is because we only tested the system on colonoscopy, we cannot yet confirm the system's generalizability. The next step is to test this system on other preventive orders, and evaluate performance. Similarly, the data may also be overfitted to the VUMC's nuances and specific clinical variations. The next step is to use data from other medical centers as input. Another limitation is that though we attempted to account for non-preventive colonoscopies by making a subpopulation with gastro-related departments removed, it is likely that there are still some colonoscopies which were ordered as a diagnostic from non-gastro departments for other reasons, such as rectal bleeding. The last limitation of note is that the system finds guidelines from existing clinician ordering practices. If clinicians have low adherence to the guidelines or new guidelines are released, then the clinicians must change their practices accordingly without decision support. One possible solution to speed up the classifier's adaptation to new practices is to weigh more recent entries higher.

Conclusion

This study examines the problem of automatically generating recommendations for preventive orders using current ordering practices. The system differs from previous work in that it uses machine learning to predict if an order should be placed instead of groups of orders, and incorporates operational features to predict if a screening is needed at the time. The system outputs a classifier for colonoscopy that achieved an AUC of 0.893. This classifier achieved an equal sensitivity to a simple model based on existing guidelines, and a higher specificity. This work demonstrates the feasibility of utilization of EMR data to learn preventive order placement patterns, which can be used as recommendations in CPOE systems. We plan to expand this system to all preventive orders.

Acknowledgements

This research was supported by training grant T15LM007450 and CTSA award No. UL1TR000445.

References

1. Katz DL, Ali A. Preventive medicine, integrative medicine and the health of the public. Commissioned paper for Institute of Medicine of the National Academies. Summit on Integrative Medicine and the Health of the Public; 2009 Feb 25-7.
2. Brawley O, Byers T, Chen A, et al. New American Cancer Society Process for Creating Trustworthy Cancer Screening Guidelines. J Am Med Assoc. 2011; 306(22):2495-9.
3. Pham HH, Schrag D, Hargraves L, Bach PB. Delivery of preventive services to older adults by primary care physicians. J Am Med Assoc. 2005; 294(4): 473-81.
4. American Cancer Society. Colorectal Cancer Facts & Figures 2014-2016. Atlanta: American Cancer Society, 2014.
5. Farley TA, Dalal MA, Mostashart F, Frieden TR. Deaths preventable in the U.S. by improvements in use of clinical preventive services. Am J Prev Med. 2010; 38(6): 600-9.
6. Jaspers MW, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. J Am Med Inform Assoc. May 2011; 18: 327-34.
7. Waitman LR, Miller RA. Pragmatics of implementing guidelines on the front lines. J Am Med Inform Assoc. 2004; 11(5):436-8.
8. Osheroff JA, Teich JM, Middleton B, Steen EB, Wright A, Detmer DE. A roadmap for national action on clinical decision support. J Am Med Inform Assoc. 2007; 14:141–5.
9. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. American Association for the Advancement of Artificial Intelligence. 1996; 37-54.
10. 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.
11. Klann J, Schadow G, McCoy JM. A recommendation algorithm for automating corollary order generation; AMIA Annu Symp Proc. 2009; pp. 333–7.
12. 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: 819-23.
13. Chen JH, Altman RB. Automated physician order recommendations and outcome predictions by data-mining electronic medical records. Paper presented at the American Medical Informatics Association 2014. Stanford University: Stanford, CA.
14. Giuse DA. Supporting communication in an integrated patient record system. AMIA Annu Symp Proc. 2003. p. 1065.
15. U.S. Preventive Services Task Force. Screening for colorectal cancer: recommendation and rationale. Ann Intern Med. 2002; 137(2): 129-31.
16. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics. 2007; 8:25.