A Recommendation Algorithm for Automating Corollary Order Generation

Jeffrey Klann[1,2], Gunther Schadow, MD, PhD[1,2], JM McCoy, MD[1,2]
[1]Regenstrief Institute, Indianapolis, IN; [2]Indiana University, Indianapolis, IN

Abstract

Manual development and maintenance of decision support content is time-consuming and expensive. We explore recommendation algorithms, e-commerce data-mining tools that use collective order history to suggest purchases, to assist with this. In particular, previous work shows corollary order suggestions are amenable to automated data-mining techniques. Here, an item-based collaborative filtering algorithm augmented with association rule interestingness measures mined suggestions from 866,445 orders made in an inpatient hospital in 2007, generating 584 potential corollary orders. Our expert physician panel evaluated the top 92 and agreed 75.3% were clinically meaningful. Also, at least one felt 47.9% would be directly relevant in guideline development. This automated generation of a rough-cut of corollary orders confirms prior indications about automated tools in building decision support content. It is an important step toward computerized augmentation to decision support development, which could increase development efficiency and content quality while automatically capturing local standards.

Introduction

There is growing evidence that decision support systems (DSSs), which assist physicians with helpful, computer-based reminders, can improve care [1]. Such systems often rely on carefully-constructed, hand-crafted guidelines to generate their reminders (e.g., [2]). However, manual development and maintenance of decision support content is expensive and time-consuming [3]. As the cost of medicine continues to rise rapidly [4], minimizing expensive human labor becomes increasingly important. Therefore, this paper explores bringing recommendation algorithms to bear on the expensive problem of manual content development.

Recommendation algorithms, prevalent in e-commerce, suggest items to customers by combining their purchase history and preferences with those of other customers [5]. Such algorithms take many forms, but the best known may be Amazon.com, which personalizes the entire shopping experience for each user [6]. Such personalization could also benefit physicians, assisting them in choosing the best tests, procedures, and medications for their patients. This is the heart of both personalized and evidence-based medicine: combining many sources of external health information with clinician expertise to provide each patient the best possible care [7, 8].

There has already been interest in using data mining to assist medical content development [9, 10]. One study targeted development of a particular type of decision support content, the corollary order [10]. We hypothesized we could develop an algorithm to generate suggestions suitable for corollary orders, with enough accuracy that the results could be quickly analyzed by a physician.

Corollary Orders

Corollary orders are trigger and response pairs that cause DSSs to suggest consequent orders in response to an antecedent order. (An example is `warfarin→prothrombin time each morning`, or, "Since you ordered warfarin, you might also be interested in ordering prothrombin time each morning.") These simple reminders have been shown to significantly improve care. One study implemented corollary orders in response to 87 trigger orders and demonstrated more than a doubling of physician compliance in a six-month inpatient trial [11].

Corollary orders are good targets for automated content generation, because their `A→B` structure is essentially the same as recommendation algorithms ("Customers who bought A also bought B").

Automated Rule Development

In a previous study, association rule mining was explored as a platform for automating corollary order development [10]. That study used the best-known association rule mining algorithm (Apriori) to find ordering suggestions [12]. However, Apriori is optimized for finding large itsemsets (such as the order sets found in previous studies [9, 10]). Hence this algorithm is very often used in retail, to understand customer purchasing patterns [5]. However, because corollary orders are similar to e-commerce product recommendations, we felt that adapting a recommendation algorithm specifically designed for that purpose is a better choice for finding corollary order suggestions.

We developed an item-based collaborative filtering algorithm, the type used by sites like Amazon.com [6], combined with "interestingness measures" (statistics relating occurrence and co-occurrence) which are used to detect association strength [13]. E-commerce algorithms like Amazon's find co-occurences of item

```
ChoosePairs():
For each distinct possible order, I
 For each diagnosis, D, in which I was ordered
  For each other order, J, used in treating D
   Increment rule I->J's frequency by 1

FilterPairs():
Given thresholds t,a,b
For each chosen pair (A,B):
  Remove if leverage(A,B)<t
  Remove if count(A)>a or count(B)>b
  Sort the remainder by conviction(A,B)
```

Figure 1: Our collaborative filtering algorithm. *ChoosePairs()*: aggregate co-occurrence of orders. *FilterPairs()*: choose top ordering suggestions using the output from ChoosePairs().

pairs among customers and compute the pairs' similarities. Our adaptation finds co-occurences of order pairs used in treating a diagnosis and computes the pairs' interestingness[1]. The method for building suggestion pairs is shown in ChoosePairs() in Figure 1. Note that in ChoosePairs(), diagnosis is merely an aggregation variable used to capture common ordering patterns across all diagnoses and is never used explicitly.

After suggestion pairs have been constructed, interestingness measures (seen in Table 1) are used to choose the best suggestions. The techniques used here, which choose the best-supported novel rules, are shown in FilterPairs() in Figure 1. First, pairs that are not novel enough (low leverage) are removed. Next, pairs that involve a somewhat common antecedent or an extremely common consequent are removed, because such suggestions are likely not to be accurate (e.g., a Complete Blood Count can precede many orders but is not predictive). Finally, remaining pairs are sorted by association strength (conviction). This is similar to the standard support-confidence approach [14], but measures like these have been shown to produce better results for most applications [13].

Methods

We tested our hypothesis on CPOE (Computer Physician Order Entry) data from an Indianapolis hospital. We presented our top results to an expert panel of two practicing internists, who counted both clinically meaningful suggestions and those they found directly relevant in guideline development, and they explained their choices to us. Here we present our algorithm and an evaluation of the top results.

Our data consisted of 2762 order types used to treat 1816 primary discharge diagnoses at the Wishard Memorial Hospital in Indianapolis. The dataset contained all 866,445 de-identified inpatient orders entered into the CPOE system in 2007. This retrospective analysis was approved by the IRB (EX0811-29).

[1] Similarity measurements are not suitable to the task of corollary ordering. A good corollary for order for A is less likely to be a *similar* to B than a *frequently co-occuring* with C (because it would therefore be useful in conjunction with A).

Classic Measures	
support(A,B)	Proportion of transactions containing A,B.
confidence(A→B)	Positive predictive value of A→B.
Our Measures	
leverage(A,B)	The proportion of transactions containing both A and B minus what would be expected if A and B were statistically independent. (Measuring information gain.)[15]
conviction(A→B)	Ratio of how frequently A appears without B over what would be expected if they were statistically independent. (Measuring implication.)[16]

Table 1: Rule interestingness measures. Top: classic filtering and sorting measures, respectively. Bottom: Similar but better-performing measures used here.

To increase support for medication suggestions, individual medication orders were first classified into medication categories. This was done using the local terminology dictionary within Wishard's CPOE system which contains medication sets and synonyms. 540 medication order types were classified into 134 categories, affecting 211,650 orders. Next, the algorithm in Figure 1 was applied, which we implemented as a SQL function. ChoosePairs() created 147,124 pairs. We set the leverage threshold (t) to 5%, which was chosen by starting at a threshold of 0% (statistical independence) and gradually increasing until relevant suggestions were truncated. The antecedent and consequent filtering thresholds (a,b) were set to remove the top 147 and 85 orders, respectively, which upon careful inspection of the top 500 orders, were minimal values for removal of most non-predictive common orders.

After FilterPairs(), 1694 suggestion pairs remained. Further filters were applied to the rules, removing rule types we were not interested in: a) non-clinical orders (such as 'egg-crate mattress' and 'pull chart'), b) ventilator orders (which appeared non-predictive in early evaluation), and c) pairs of cerebrospinal fluid and urine tests (which were elements of common panels and never ordered separately). This left 584 rules.

A test set was created using the 92 top suggestions. These were combined with twenty randomly chosen poor suggestions (leverage$<$0) as a control, so as to not bias the raters. All 112 suggestions were ordered randomly. This test set was presented to our expert panel of two Board Certified Internists, who were asked to: a) choose which suggestions were clinically correct (A and B are clinically associated), b) from among the

correct suggestions, choose which they felt were directly relevant in guideline development, and c) explain their choices. We examined their ratings and measured the inter-rater agreement with Cohen's κ coefficient. Then we analyzed the types of suggestions generated, and we compared the correctness rating to conviction.

Because the inter-rater agreement of correctness was moderate-to-strong, we produced an adjusted set that included only suggestions on which the raters agreed (presuming disagreement indicated an error). There was no statistical agreement in our subjective relevance measure, which was expected since the rating was based on personal experience. However, the union of relevance ratings is a reasonable indication of potential overall suggestion value, so we produced an adjusted relevance set of the union of agreed relevance on the adjusted correctness set.

Results

Table 2 shows the findings of the expert panel.

Score	Rater 1	Rater 2	(κ) Adjusted
$\frac{Correct}{Total}$	71.7%	70.6%	(.603) 75.3%
$\frac{Relevant}{Total}$	44.8%	33.6%	(-.225) 47.9%

Table 2: Rater scores of top 92 suggestions for clinical correctness and relevance in guideline development.

The raters agreed that 21 suggestions were both correct and relevant, some of which can be found in Table 3. Analyzing these revealed two types of suggestions: those which applied when treating any patient, and those that made sense only when treating or considering certain diagnoses. There was agreement on which were in each category.

Routine Care	
Hgb A1c→Diabetic Diet	Diabetes
Retic Count→Type&Cross	Anemia
Occ.Therapy→Phys.Therapy	Routine
Folate Level→B12 Level	Alcoholism
B-Natriuretic Peptide→EKG	Heart Failure
Warfarin→Occult Blood Test	Blood Thinners
Feeding Tube→Abdomen XR	Feeding Tube

Specific Situation	
B12 Level→TSH	Dementia
Abdomen Ultrasound→Lipase	Pancreatitis
B-Natriuretic Peptide→TSH	Heart failure
Metoclopramide→Abdomen XR	Nausea
Hgb A1c→Lipid Profile	Diabetes

Table 3: Suggestions agreed to be both correct and relevant (n=21). Top: universally applicable. Bottom: specific situations.

Rater disagreement regarding correctness occurred in 14 cases, but only two appeared to be true errors caught by only one rater. The other 12 were disagreements over what rules were too weak or incidental to be correct. Examining relevance disagreements was also informative. Although the 16 disagreements in the adjusted set were largely opinion, 25% of the time one rater indicated a suggestion was part of a panel while the other didn't. Examples of disagreements are listed in Table 4.

Correctness Disagreements	
Insulin→Narcotics	Diabetic ketoacidosis
Hgb A1c→TSH	Initial diabetes workup
Relevance Disagreements	
Trapeze→Neurovascular Checks	Neurology patients
Osmolality→Urine Lytes	Acid/base imbalance

Table 4: Suggestions rated correct or relevant by only one rater.

Analyzing suggestions rated incorrect or irrelevant, we found they fell into six categories, shown in Table 5, with examples. There was virtually no disagreement over the incorrectness of the Inverted or Terminology types, though those only accounted for 15.8% of the agreement set of incorrect suggestions. The majority of incorrect suggestions were the Weak or Incidental type (and often were so weak that there was no discernible connection). The correct but irrelevant suggestions were fairly evenly distributed between the two types.

Incorrect Suggestion Types	
Incidental	Routinely co-occuring in specific situations. ICU K Replacement→Neurovitals Arterial Line→H2 Antagonists
Inverted	The consequent implies the antecedent. Gentamicin Level→Gentamicin Inj. Routine Line Care→PICC Insertion
Weak	The association is too specific or too unclear Midazolams→Type&Cross Sputum Culture→Vancomycin Level
Terminology	A and B mean the same thing. Haloperidol Tab→Haldol Med
Irrelevant Suggestion Types	
Panel Order	A and B are automatically ordered together. T3-Free Lvl→T4-Free Lvl
Obvious	A physician would never order A without B. Abdomen CT→Pelvis CT

Table 5: Types and examples of suggestions rated as incorrect or irrelevant by both raters.

Finally, we examined the accuracy of the conviction-based sort by plotting adjusted correctness against conviction, in groups of five suggestions. This can be seen in Figure 2. While there is a downward trend, the trend is not monotonic. Therefore, conviction worked better than random sorting, but it was not extremely predictive of correctness.

Discussion

Our recommendation algorithm finds clinically relevant corollary orders based entirely on automated data

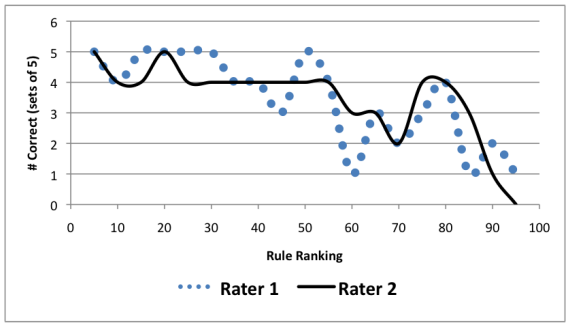

Figure 2: Rated correctness of top 92 rules when sorted by conviction.

mining of *a posteriori* treatment data, creating a list of potential corollary orders that are clinically meaningful 75.3% of the time and directly relevant 47.9% of the time. This does not replace human reasoning, but it does allow a rater to quickly choose the best suggestions. This is already more efficient than completely manual development, if only because eliminating 'chaff' is simpler than developing 'wheat' from scratch. Additionally, the results are already encoded and reflect local practice patterns[10] that can be combined with other evidence-based resources.

The correct and relevant corollary orders could be utilized in several ways. Top-rated suggestions could be reminder-style corollary orders, with more specific suggestions appearing only when treating appropriate diagnoses. To reduce reminder fatigue [17], the lower-rated (but still relevant) orders could be less obtrusively placed, perhaps by listing them in a non-modal 'suggested orders' box in the CPOE system.

Suggestions not relevant as corollary orders could still be helpful in other contexts. For example, 1. the strength of obvious associations could be monitored to measure quality of care in the hospital, 2. weak suggestions could be investigated to see if they correlate with poor practice standards (e.g., we saw `Colon Lyte Lavage→Bisacodyl`, which probably shouldn't occur together), and 3. the frequency of some non-orders could be used to generate helpful statistics of patient types seen in the hospital (e.g., orders for `pull chart` might indicate a complex patient or `egg-crate mattress` might indicate a patient at risk of an extended stay or skin breakdown).

Rater disagreements often reflected opinion, highlighting that raters are not themselves automatically tailored to local practice patterns. What is too obvious or specific for a corollary order will vary by practice, and definitions of common panels will vary by CPOE system. Therefore in an actual implementation, it becomes important to choose a rater familiar with the care standards of the target facility. Furthermore, future evaluations would benefit from an objective relevance standard, such as cross-referencing with guidelines published by the National Guideline Clearinghouse [18].

Limitations and Future Directions

Not surprisingly, issues with data and terminology hampered the system. For example, although we realized that automatic removal of common panel orders would increase relevant results, this data was not available. Instead we had "order components," which contained some (but not all) panels, and it also included categorical groupings like `Diets` and `Vaccines`. An order component filter removed 11 suggestions, only 7 of which were panel orders. The other 5 were potentially relevant suggestions (e.g., an alternate antihistamine), so we removed this filter.

Further, even with Wishard's extensive terminology dictionary of 22,700 terms, 14,492 synonyms, and 4070 medication categories, terminology-related reasoning problems were still an issue. For instance, the test set contained `Haloperidol Tab→Haldol Med`, because the Haldol Medications set did not include Haloperidol tablets. Also, there was no way to categorically exclude non-order items. Because Wishard's terminology dictionary is more complete than most, we expect the algorithm to have more terminology problems in other environments. One way to combat this would be through statistical measures of synonomy and relationship, using concept matching tools like those in the UMLS.

Also, since Wishard is a teaching hospital, our dataset is probably noisy, because residents may be more likely than experienced physicians to order unnecessary tests. One approach to noise removal would be filtering orders from inexperienced physicians, as is done with outlying customers by Amazon.com's algorithm [6].

Sometimes the data itself is incorrect. `Routine Line Care→PICC Insertion` is an erroneous suggestion that could be filtered by the right heuristic, utilizing the fact that a line care order should only (but not always) occur with a PICC order. However, in our data, `Routine Line Care` was ordered without a PICC 26 times!

The algorithm's design itself also led to occasional unexpected results. For example, `Gentamicin Level→Gentamicin Injection` appeared because the antecedent/consequent thresholds filtered out the correct (inverted) rule. This leads us to believe that simple antecedent and consequents thresholds are not the best solution.

Further improvements could make the algorithm better able to find the "sweet spot" that is neither too obvious nor too specific. For one, although sorting by

conviction is better than randomizing, Figure 2 indicates better options might exist. Also, although our informal tests showed that a leverage-based threshold is superior to common interestingness thresholds (such as support), perhaps frequency-based interestingness measures are not the best choice at all. Although they succeed in e-commerce market basket analyses [19], there the goal is only to sell products (and not necessarily even the best products). Medical decision support is far more nuanced, can seriously affect patients' lives, and is constantly in danger of being ignored by busy doctors [3]. Increasing the complexity of the reasoning (for example, by including the temporal relationships of A and B or using supervised learning as feedback into the algorithm) might be promising here.

Finally, to produce suggestions for an evidence-based guideline, there must be some way of excluding poor practice standards and treatments that did not work or that the target facility does not recommend. Therefore, although computer methods involving deeper data analysis could eventually aid in some of this (e.g., by analyzing patient readmissions), it also highlights the importance of having humans in the loop.

Conclusions

This paper has demonstrated an automated technique using data mining to produce a 75.3%-accurate roughcut of a corollary orders list. After initial configuration, the technique can quickly generate such lists without human intervention and with no additional overhead, using entirely algorithmic analysis of existing data. Such lists could be given to expert panels for guidance in implementation of corollary order protocols or could be combined with other resources to create evidence-based decision support content automatically tailored to local practice standards.

This study is one step toward mature, automated tools which will assist development of decision support content for CPOE systems. Such tools could boost efficiency in decision support content development. Additionally, they could personalize the medical ordering experience in the same vein that recommendation algorithms automatically customize Amazon.com for each user. Finally, these tools would also have important applications for quality improvement and care management.

Acknowledgements

Thanks to Marc Overhage, Peter Szolovits, Mike Barnes, Joe Kesterson, Jon Duke, and Kevin Chang. This work was performed at the Regenstrief Institute, Indianapolis, IN, and was supported in part by grant T15 LM07117 from the National Library of Medicine.

References

[1] 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):1409–1416.

[2] Litzelman D, Dittus R, Miller M, Tierney W. Requiring physicians to respond to computerized reminders improves their compliance with preventive care protocols. Journal of General Internal Medicine. 1993 Jun;8(6):311–317.

[3] Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. Journal of the American Medical Informatics Association. 2003;10(6):523–530.

[4] Office of the Actuary: Centers for Medicare & Medicaid Services. National Health Expenditure Projections 2007-2017. 7500 Security Boulevard, Baltimore, MD 21244: Dept. Health & Human Services; 2006.

[5] Schafer JB, Konstan JA, Riedl J. E-Commerce Recommendation Applications. Data Mining and Knowledge Discovery. 2001;5(1):115–153.

[6] Linden G, Smith B, York J. Amazon. com Recommendations: Item-to-Item Collaborative Filtering. IEEE INTERNET COMPUTING. 2003;p. 76–80.

[7] Fierz W. Challenge of personalized health care: to what extent is medicine already individualized and what are the future trends. Med Sci Monit. 2004;10(5):123.

[8] Sackett DL. Evidence-based medicine. Seminars in Perinatology. 1997 Feb;21(1):3–5.

[9] Santangelo J, Rogers P, Buskirk J, Mekhjian HS, Liu J, Kamal J. Using data mining tools to discover novel clinical laboratory test batteries. AMIA Annual Symposium Proceedings / AMIA Symposium AMIA Symposium. 2007;p. 1101. PMID: 18694198.

[10] Wright A, Sittig DF. Automated development of order sets and corollary orders by data mining in an ambulatory computerized physician order entry system. AMIA Annual Symposium Proceedings / AMIA Symposium AMIA Symposium. 2006;p. 819–23. PMID: 17238455.

[11] Overhage JM, Tierney WM, Zhou XA, McDonald CJ. A Randomized Trial of "Corollary Orders" to Prevent Errors of Omission. Journal of the American Medical Informatics Association. 1997 Oct;4(5):364–375. PMC61254.

[12] Agrawal R, Srikant R, Bocca JB, Jarke M, Zaniolo C. Fast Algorithms for Mining Association Rules. In: Proc. 20th Int. Conf. Very Large Data Bases,VLDB. Morgan Kaufmann; 1994. p. 487–499.

[13] Lallich S, Teytaud O, Prudhomme E. III.3. In: Association Rule Interestingness: Measure and Statistical Validation. Springer; 2007. p. 251–275.

[14] Bramer M. Principles of Data Mining. Springer; 2007.

[15] Piatetsky-Shapiro G. Discovery, Analysis, and Presentation of Strong Rules. In: Piatetsky-Shapiro G, Frawley W, editors. Knowledge Discovery in Databases. Cambridge, MA: AAAI/MIT Press; 1991. p. 229–248.

[16] Brin S, Motwani R, Ullman JD, Tsur S. Dynamic itemset counting and implication rules for market basket data. In: Proceedings of the 1997 ACM SIGMOD international conference on Management of data. Tucson, Arizona, United States: ACM; 1997. p. 255–264.

[17] Payne TH, Nichol WP, Hoey P, Savarino J. Characteristics and override rates of order checks in a practitioner order entry system. Proceedings of the AMIA Symposium. 2002;p. 602–606. PMC2244252.

[18] National Guideline Clearinghouse. NGC; 2009. http://www.guideline.gov/.

[19] Chen L, Sakaguchi T. Data Mining Methods, Applications, and Tools. Information Systems Management. 2000;17(1):65.