Journal of Biomedical Informatics 43 (2010) 891–901

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

An automated technique for identifying associations between medications, laboratory results and problems

Adam Wright [a,b,*], Elizabeth S. Chen [c,d], Francine L. Maloney [e]

[a] Division of General Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[b] Clinical and Quality Analysis, Partners HealthCare, Boston, MA, USA
[c] Center for Clinical and Translational Science, University of Vermont, Burlington, VT, USA
[d] Division of General Internal Medicine, University of Vermont College of Medicine, Burlington, VT, USA
[e] Clinical Informatics Research and Development, Partners HealthCare, Boston, MA, USA

ARTICLE INFO

Article history:
Received 27 October 2009
Available online 25 September 2010

Keywords:
Data mining
Association rule mining
Clinical decision support

ABSTRACT

Background: The patient problem list is an important component of clinical medicine. The problem list enables decision support and quality measurement, and evidence suggests that patients with accurate and complete problem lists may have better outcomes. However, the problem list is often incomplete.
Objective: To determine whether association rule mining, a data mining technique, has utility for identifying associations between medications, laboratory results and problems. Such associations may be useful for identifying probable gaps in the problem list.
Design: Association rule mining was performed on structured electronic health record data for a sample of 100,000 patients receiving care at the Brigham and Women's Hospital, Boston, MA. The dataset included 272,749 coded problems, 442,658 medications and 11,801,068 laboratory results.
Measurements: Candidate medication-problem and laboratory-problem associations were generated using support, confidence, chi square, interest, and conviction statistics. High-scoring candidate pairs were compared to a gold standard: the Lexi-Comp drug reference database for medications and Mosby's Diagnostic and Laboratory Test Reference for laboratory results.
Results: We were able to successfully identify a large number of clinically accurate associations. A high proportion of high-scoring associations were adjudged clinically accurate when evaluated against the gold standard (89.2% for medications with the best-performing statistic, chi square, and 55.6% for laboratory results using interest).
Conclusion: Association rule mining appears to be a useful tool for identifying clinically accurate associations between medications, laboratory results and problems and has several important advantages over alternative knowledge-based approaches.

1. Introduction

Many applications in biomedical informatics require clinical knowledge bases, e.g. relating signs and symptoms to diseases (for automated diagnosis), screening tests and indications (for preventive care decision support) or diseases and medications (for indication-based prescribing). These knowledge bases are often developed and maintained by experts, at significant cost. However, automated methods (usually statistical) for developing such knowledge bases hold promise. In this paper, we describe a set of data mining techniques which can be used to automatically infer (and measure the strength of) relationships between medications, laboratory results and problems, and validate a knowledge base we developed using the technique against two gold standards. We also describe a particular potential application of this knowledge base: closing "gaps" in patient problem lists.

* Corresponding author at: Division of General Medicine and Primary Care, Brigham and Women's Hospital and Harvard Medical School, 1620 Tremont Street, 3rd floor, Boston, MA 02120, USA. Tel.: +781 416 8764; fax: +617 732 7072.
E-mail address: awright5@partners.org (A. Wright).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.09.009

2. Background

2.1. Clinical problem lists

Electronic and paper medical records have long been organized into a variety of sections such as visit notes, medication lists, laboratory results and problem lists. The problem list has been a standard part of the medical record for a considerable period of time, but it began to occupy a central place in the diagnostic reasoning process with Larry Weed's seminal 1968 paper "Medical Records that Guide and Teach", which introduced the concept of the problem-orient medical record (POMR) [1] and the ability to create and

maintain a structured, coded problem list is today a requirement of the Certification Commission for Health Information Technology (CCHIT) for all certified electronic health record (EHR) systems [2]

The problem list is important for a variety of reasons. First, having an accurate problem list enables clinicians to see the full spectrum of a patient's problems and is a key input to diagnostic reasoning. The problem list is also a communication tool – when a physician is seeing a patient for the first time (perhaps in a consultation or while providing coverage for the patient's regular physician), the problem list allows him or her to understand the patient's issues. Evidence suggests that patients with complete problems lists may receive higher quality care than patients with gaps in their problem list [3].

In addition to the apparent direct clinical benefits of a complete problem list, the problem list has a number of important ancillary benefits. Clinical decision support systems often depend on coded data elements, and one study found that 22.3% of decision support rules at Partners HealthCare depended on problems [4]. An accurate, complete, coded problem list is also critical for quality measurement and research.

Although the problem list is important, there is also substantial evidence to suggest that it is often woefully incomplete. A study by Szeto et al. found that among patients with coronary artery disease, only 49% had the problem on their problem list (accuracy ranged from 42% for benign prostate hypertrophy to 81% for diabetes) [5]. Szeto also studied the specificity of problems, and found that it was extremely high: 98–100% of patients with a problem on their list actually had the problem, suggesting that false positives are low.

2.2. Inferring clinical problems

Because gaps in the problem list are common, researchers have explored methods for automatically inferring problems. These efforts have principally fallen into two categories: proxy methods and natural language processing (NLP)-based methods. Proxy methods attempt to use other clinical data in the EHR to infer problems. Burton and Simonitas described an approach for inferring problems from medications in EHRs based on drug indication data found in the NDF-RT, a standard drug terminology [6]. Carpenter described a similar system which used expert-developed rules to locate potential drug-problem mismatches in diabetes [7]. Lin and Haug described a more sophisticated system based on Bayesian networks [8]; the Lin system focused on five specific diagnoses and used a knowledge base of clinical variable-diagnosis associations derived from internal medicine text books, but tuned the algorithm using Bayesian networks.

Proxy methods that use administrative (e.g., claims) data have also been proposed. Poissant et al. used medication claims from a Canadian provincial insurance system coupled with an expert-developed knowledge base of single-indication drugs to identify problem list gaps with some good success [9].

In addition to proxy methods, a variety of NLP-based methods have also been proposed. Such systems extract candidate problems from unstructured clinical text, such as progress notes. Meystre and Haug described an NLP-based system, using the National Library of Medicine's MetaMap Transfer application [10] and NegEx [11] for negation detection that focused on 80 specific medical problems [12,13]. They reported an initial precision of 0.756 and recall of 0.740. By customizing the MetaMap dictionary, they were able to increase their recall to 0.896 with only a slight recall tradeoff. Similar systems have been described using other NLP engines [14].

2.3. Developing a knowledge base using automated techniques

Building on this prior work, we explored and developed data mining techniques to automatically identify associations between problems and structured non-problem data in the EHR (medications and lab results in this analysis). Our goal was to develop a knowledge base of medication-problem and laboratory result-problem associations in an automated fashion using data mining techniques, and to evaluate it. This knowledge base would have a variety of applications, foremost among them inferring clinical problems.

The data mining techniques that we used, frequent item set mining and association rule mining [15] are not themselves novel. The techniques have been developed in computer science for over a decade and have been used in a variety of fields [16–18]. Association rule mining underpins Amazon's recommendation feature, which suggests books based on the tastes of others whose past purchase history is similar to yours [19]. We describe frequent item set mining and association rule mining, and our extensions to them, in detail in Section 3.

Association rule mining and related techniques have been used previously in medical informatics. Cao et al. used NLP and co-occurrence statistics to discover disease-finding co-occurrences in discharge summaries with strong results [20]. Wang et al. used similar techniques to locate potentially unknown adverse effects of drugs [21], Mullins et al. used the techniques for public health surveillance [22] and several other applications have also been reported in the literature [23,24]. The authors of this paper have also previously used association rule mining to analyze clinical information system log files [25], locate disease-drug associations in the biomedical literature and clinical text [26] and develop order sets and corollary orders in an automated fashion [27].

2.4. Hypothesis

We hypothesize that association rule mining will be a feasible technique for analyzing a large clinical data set to successfully identify clinically accurate and meaningful associations between structured data elements (specifically medications and laboratory results) and problems in the EHR. We also hypothesize that the cost of generating such associations through data mining will be less than through manual knowledge base creation and that the volume of rules generated will outstrip existing expert-curated knowledge bases. Finally, we hypothesize that the empiric basis and inherent measurability of the rules developed through these techniques will allow them to be more readily characterized and validated than expert-derived rules without a similar empiric basis.

3. Methods

In our project, we used two related data mining techniques: frequent item set mining and association rule mining. Frequent item set mining is a technique for locating commonly co-occurring items in a transaction database. Association rule mining is an extension of frequent item set mining, which looks at the direction of association in addition to simple co-occurrence. The two techniques are closely related and complementary; in fact, the output of frequent item set mining algorithms can be used as the input to many association rule mining algorithms.

3.1. Frequent item set mining

Frequent item set mining, as introduced in the background Section 2, is an important tool for assessing the co-occurrence of items in a transactional database and, thus, for determining possible associations among them. To describe the technique, we must introduce some formalism. We begin with the set I, which contains all of the items which might appear in the transaction database. In

a grocery store example, I would contain all of the items available for purchase in the store; in a clinical example, I might contain all of the medications, procedures and laboratory tests which might be orderable in a hospital. The next concept is the transaction T_i. Each transaction is a set of items that occur together in some logical grouping which we call a transaction. In the grocery example, a transaction might be all the items purchased together by a customer (e.g., the contents of their basket, which is why this frequent item set mining is sometimes called "market basket analysis"). In the clinical example, it might be all the orders for a patient in a particular admission (or perhaps longitudinally). Each $Ti \subseteq I$. We refer to a database D, which contains all of the transactions $T_0 \ldots T_n$.

The concepts introduced thus far (item, transaction and database) are needed to characterize what happened in a particular transactional setting. However, with frequent item set mining, our goal is to determine which items in I naturally occur together within the database D. We call these co-occurring groups of items item sets, and we term candidate item sets X. We define the cover of a candidate item set X to be the set of transactions in D that contain X. The support of X is defined as the number of items in the cover of X (i.e., support(X) = |cover(X)|). We should note that sometimes support is alternatively defined as |cover(X)|/|D|, or the proportion of transactions in D that contain X.

3.2. An example

To illustrate these concepts, it is helpful to introduce an example. Consider five patients whom we will characterize only by their medications and problems. These patients are shown in Fig. 1.

In this example, the set I consists of all the unique medications and problems, and contains I = {diabetes, hypertension, insulin, lisinopril, metformin, multivitamin, polycystic ovarian syndrome}. Each patient can be thought of as a transaction T_i in the clinical database D. One can readily observe some frequent item sets that appear to be promising. For example, the item set X = {lisinopril, hypertension} has support of 2 (since two patients have both lisinopril and hypertension in their set transaction). The cover of X is {pt 1, pt 2}. The item set {metformin, diabetes} and {insulin, diabetes} also have support of 2. Patient 5 provides a possible counterexample to this apparent association between metformin and diabetes, however. The statistical measures we use to account for this will be described later.

3.3. An efficient algorithm

It seems, from this example at least, that frequent item set mining may be a useful technique for determining the relationships between data elements. In the simple example above, one can mentally identify the possible frequent item sets and compute their support. However, when the number of items is high, frequent item set mining poses a substantial computational challenge. Indeed, given a set of items I, there are $2^{|I|}$ candidate item sets. For a small item set, this computation may be tractable. However, it quickly becomes intractable when the size of the item set is large. For example, a two year sample of data from the Brigham and Women's Hospital (BWH) shows a total of 25,848 unique data elements recorded across the medication, laboratory result and problem domains (this includes only coded elements – when uncoded data elements are included the number is much higher). This means the total candidate item set space contains $2^{25848} = 1.055 \times 10^{7781}$ members, which is computationally intractable.

Pt 1: {lisinopril, multivitamin, hypertension}

Pt 2: {insulin, metformin, lisinopril, diabetes, hypertension}

Pt 3: {insulin, diabetes}

Pt 4: {metformin, diabetes}

Pt 5: {metformin, polycystic ovarian syndrome}

Fig. 1. The medications and problems for 5 sample patients.

In 1993, Rakesh Agrawal, of IBM's Almaden Research Center, described an efficient algorithm for computing the complete set of frequent item sets with support greater than a minimum threshold from a database [28]. This algorithm exploits a property of the support metric first described by Agrawal: downward closure. This property states that:

$$X \subseteq Y \Rightarrow \text{support}(Y) \leq \text{support}(X)$$

the property follows trivially from the fact that:

$$\text{cover}(Y) \subseteq \text{cover}(X)$$

In words, it means that, given a candidate item set X, any item set Y which fully contains X must have support less than or equal to the support of X. In other words, if you extend the item set X by adding an item to it, the support must either remain the same or go down, it cannot increase.

The Apriori algorithm has four phases: initiation, joining, pruning and evaluation. The algorithm begins with an initiation phase. In this phase, all 1-item sets with support > minimum support (minsup) are generated. These 1-item sets are frequent (because their support > minsup) so they are added to the result. Next, all of these 1-item sets are joined with each other to produce 2-item sets (the joining phase). Each of these 2-item sets is evaluated to determine whether the item set exceeds the support threshold (the evaluation phase). If it does, it is added to the result. Next, the 2-item sets are combined to form 3-item sets. However, from this point, another step is added: the pruning phase. Each 3-item set is checked to see whether it contains any 2-item sets or 1-item sets that are not in the result set (e.g., non-frequent). The upper bound of the support for any particular 3-item set is, by the downward closure property of support, the support of its least frequent subset. If any subset of the 3-item set is not frequent (e.g., not in the result set), then the 3-item set itself must be non-frequent and does not need to be evaluated. If all subsets of the 3-item set are frequent, then we proceed to the evaluation phase.

Although a naïve algorithm that performed only the initiation, joining and evaluation steps would work correctly, the pruning step has two key advantages over such a naïve algorithm. First, the evaluation space is much smaller, since many candidate item sets are excluded during pruning (and, additionally, all supersets of excluded item sets are also pruned, further limiting the search space). Second, the algorithm naturally terminates when there can be no further joining of k-item sets into k + 1-item sets that are not all pruned. This allows the algorithm to terminate much sooner than the naïve algorithm which must process the entire power set of the item set before terminating.

3.4. Association rule mining

Frequent item sets, by themselves, are inherently nondirectional. An item set is considered frequent if its support exceeds the support threshold. However, some relationships between items may have a direction. In the example given in Fig. 1, the {insulin, diabetes} relationship is directional. This directionality is fairly obvious clinically: almost everyone who receives insulin has diabetes, but only certain people with diabetes receive insulin.

To account for this directionality, frequent item set mining is often extended to association rule mining. An association rule is an expression $X \rightarrow Y$ where X is an item set, Y is an item set and X

and *Y* are disjoint. Like item sets, association rules can be characterized by their support. We say that support($X \rightarrow Y$) = support($X \cup Y$). This measure, however, is symmetric (i.e., support(X Y) = support($Y \rightarrow X$)). In order to account for directionality, we introduce another measure: confidence. We define confidence($X \rightarrow Y$) = support($X \cup Y$)/support(X). The confidence is the proportion of all transactions containing *X* that also contain *Y*. In our example from Fig. 1, confidence({insulin} → {diabetes}) = 100% since all patients receiving insulin also have diabetes. However, confidence({diabetes} → {insulin}) = 66.7% (because one patient, pt 4, is receiving only metformin for his diabetes).

It is important to note that the directionality inferred by association rule mining is purely correlational. An implication $X \rightarrow Y$, with confidence *c*, simply means that c% of transactions containing *X* also contain *Y*. Such a relationship should not be construed as implying causation without further analysis (typically beyond association rule mining, e.g. an experiment).

3.5. Measures of Interestingness

Although association rules can be filtered by their support and confidence, there are often many more potential rules produced through these techniques than can be manually reviewed. A variety of measures of "interestingness" have been proposed which can be used to filter these item sets and association rules [29].

In this paper, we concentrate our attention on five commonly used and robust measures: support, confidence, chi square, interest (sometimes called lift) and conviction. The formulas for these statistics are given in Fig. 2. In this figure, for a given rule ($X \rightarrow Y$), *a* represents the number of transactions in the database containing both *X* and *Y*, *b* the number containing *X* but not *Y*, *c* the number containing *Y* but not *X* and *d* the number containing neither *X* nor *Y*.

Support and confidence have already been defined in this section. The chi square statistic has its usual meaning. To compute chisq($X \rightarrow Y$), one conceives of the database *D* as a two-by-two table. The upper-left cell contains the number of transactions which contain both X and Y, the upper-right the number of transactions which contain *X* but not *Y*, the lower-left the number of transactions which contain *Y* but not *X* and the lower-right the number of transactions which contain neither *X* nor *Y*. The advantage of the chi square statistic is that it accounts for the baseline frequency of *X* and *Y*. Support, on the other hand, does not: association rules may score highly simply because their members are very frequent in the database, even if the relationship between *X* and *Y* is weak.

Metric	Formula
Support	a
Confidence	$\frac{a}{a+c}$
Chi square	$\frac{(a\cdot d - b\cdot c)^2 \cdot (a+b+c+d)}{(a+b)\cdot(c+d)\cdot(b+d)\cdot(a+c)}$
Interest	$\frac{\left(\frac{a}{a+b}\right)}{\left(\frac{a+c}{a+b+c+d}\right)} = \frac{a\cdot(a+b+c+d)}{(a+b)^2}$
Conviction	$\frac{(a+c)\cdot(b+d)}{(a+b+c+d)\cdot c}$

In this figure, *a*, *b*, *c* and *d* have their usual meanings in a two-by-two table:

	Y	*Y'*
X	*a*	*b*
X'	*c*	*d*

Fig. 2. Formulation of five measures of interestingness used in the project.

Interest (or lift) is another statistic which attempts to correct for this weakness. Confidence tends to rate rules highly where the consequent (*Y*) is frequent. For example, if 80% of transactions in a database contain *Y*, then the expected confidence of any rule $X \rightarrow Y$ is 80%, even before taking the influence of *X* on *Y* into account. The interest($X \rightarrow Y$) is defined as the confidence($X \rightarrow Y$) divided by the proportion of all transactions that contain *Y*. This scales the confidence to account for the commonality (or rarity) of *Y*.

The final measure we consider is conviction, described by Brin, Motwani, Ullman and Tsur [30]. Conviction stands out among the other statistics because its derivation is actually grounded in error rates (where an error is a counter example to the rule $X \rightarrow Y$, i.e. a transaction where *X* occurs but *Y* does not). Conviction, then, is the ratio between the expected error rate assuming independence and the observed error rate. Higher values indicate greater strength of association (indeed conviction has no upper bound, and infinite conviction corresponds to the case where there were no errors observed and every transaction containing *X* also contains *Y*).

3.6. Data set

We hypothesized that association rule mining would be a useful technique for inferring relationships between medications, laboratory results and problems. Such relationships could then be used to identify potential gaps in patient problem lists. In order to explore this, we randomly selected a cohort of 100,000 patients of the Brigham and Women's Hospital. To be included in our cohort, a patient must have been seen at least once during 2007 and 2008 and have two or more outpatient notes in their record. We excluded patients who had fewer than two notes because many of them may have been seen in an acute or consultative setting and have limited documentation.

For each of these patients, we requested and received structured problems, laboratory results and medications as stored in our EHR system. The problems are coded using a proprietary problem terminology that is mapped to SNOMED CT [31]. Laboratory results are coded using LOINC [32] and the result file includes the laboratory test identifier, LOINC code, numeric result, unit of measure, text result, flags and comments. Medications are coded using a propriety medication terminology that is mapped to First Databank and also, indirectly, to RxNorm [33]. The medication file contains the medication, route and dose.

All data were de-identified and encrypted before being analyzed. Our protocol was reviewed and approved by the Partners HealthCare Human Subjects Committee.

After requesting and receiving the problem, medication and laboratory result data files, we prepared them for analysis. Problems were stripped of modifiers and qualifiers, medications were simplified to just the drug product (excluding route and dose) and laboratory results were analyzed three ways: (1) just by unique test (e.g., all CD4 tests would be viewed as identical, regardless of the result), (2) by test and flag (e.g., a high CD4 would be viewed as different from a normal or low CD4) and (3) for tests with qualitative results (which generally lack flags), by test and qualitative result (e.g., a blood smear with the result "2 + sickle cells" would be viewed as different from a smear with the result "normal morphology").

3.7. Extensions to the techniques

After a preliminary analysis, we noted two problems with conventional approaches to association rule mining that were limiting the accuracy of our results. First, we found that the associations between many anti-HIV agents and the problem HIV was lower than expected. We traced this to the fact that some of the patients had

HIV on their problem list while others had AIDS (and some had both). As a result, we developed a set of problem classes, which combined clinically related entities. These classes are described in Appendix 1.

We also found that there were some unexpectedly strong associations between apparently unrelated items which we believed were attributable to comorbidities. For example, the association rule insulin → hypertension scored highly, despite the fact that insulin is used to treat diabetes, not hypertension. However, there is strong comorbidity between the two conditions, so this rule is likely due to transitive association. We were unable to locate any method to control for this in the literature. After significant experimentation, we devised a hold-out method. In this method, significant problem-problem associations (i.e., comorbidities) are first computed. Then, whenever a candidate association is located that meets the support and confidence thresholds (we used a minimum support of 5 and confidence of 10%), we locate the comorbidities for that problem. For each comorbidity, we repeat the analysis of the candidate association is repeated on the subset of patients without the comorbid condition and evaluate the change in statistics.

Prior systems have also used transitive inference in association rule mining; however, they used it to infer additional association rules. For example, Narayanasamy et al. describe a text-mining application mining Medline for associations between diseases and genes [34]. They are interested in the situation where they locate associations A → B and B → C, but do not find an association A → C. In this circumstance, they do an additional round of evaluation to determine if A → C is also a valid association. In other words, they employ transitive association rules as a tool for generating additional candidate associations which are not otherwise found.

We use transitive inference for the reverse problem: pruning spurious candidate associations. Using our diabetes example, we found the rule insulin → hypertension. To validate this candidate association, we reviewed comorbidity data, and found a disease-disease association: diabetes → hypertension, as well as several other disease-disease associations with hypertension. We then iteratively re-evaluated the insulin → hypertension rule once for each disease comorbid with hypertension, and found that the rule fell below our threshold when diabetic patients were excluded. Based on this, we identified diabetes as a transitive mediator for the spurious insulin → hypertension rule and were thus able to automatically reduce the candidate rule set by removing the rule. This is in contrast to the Narayanasamy method, which would apply to a situation where we identified rules insulin → diabetes and diabetes → hypertension but did not identify the rule insulin → hypertension. We would apply the method, which would then propose insulin → hypertension for further study (candidate generation rather than reduction).

3.8. Analysis

Several different software packages were tested for computing association rules; however, we found that they were not adequate to perform the analyses needed. Some were unable to handle the large volume of data or required substantial reformatting of the input. Others lacked support for the statistics that we wanted to include in our analysis, and none supported (or could be easily extended to support) our novel iterative transitive reduction technique. As a result, we developed our own analysis software which implements the Apriori algorithm but is tuned specifically for clinical data. It supports all of the statistics of interest, the iterative transitive reduction technique and also provides for on-the-fly encryption and decryption of the datasets. The software was developed in C# and compiled using Microsoft Visual Studio 2008 for the .NET 3.5 Common Language Runtime. All analyses were carried out on a computer with 2 GB of memory and an Intel Core 2 Duo L7500 processor running at 1.60 GHz.

We used the software to generate the top 500 associations for drugs and labs with problems according to each of the five statistics of interest. We limited generation to rules with a single drug or lab in the antecedent set and a single problem in the consequent set in order to enable gold standard evaluation. Because there is significant overlap in the associations selected by each statistic, the total number of associations was less than 2500.

4. Evaluation

After generating the association rules for medications, laboratory results, and problems, we evaluated the rules by comparing them to a gold standard. For medications, we used the Lexi-Comp drug knowledge base (Lexi-Comp, Inc., Hudson, Ohio), which contains information including pharmacology, dosing, administration, use and contraindications of all FDA-approved drug products. For laboratory results, we used Mosby's Diagnostic and Laboratory Test Reference [35] which contains information on common laboratory test results and their uses.

For the evaluation, we identified the top 500 medication-problem and laboratory-problem associations according to each of the five statistics, yielding ten lists of 500 items (e.g., one list for the top 500 medications according to support and another for the top 500 laboratory results using chi square). We then compared each association to the reference sources to determine whether the association was also found in the gold standard reference source. Because the Lexi-Comp database contains all drugs, medication-problem associations were coded as either "indicated" or "not indicated". Mosby's Diagnostic and Laboratory Test Reference, however, did not contain some esoteric laboratory results (the number of unique laboratory tests is much larger than the number of FDA-approved medications, and we were unable to locate any gold standard which was entirely complete). Therefore, for laboratory results, each identified pair was coded as "indicated", "not indicated" or "not found".

Based on these comparisons, we computed an "accuracy" statistic – the proportion of associations found that matched the gold standard. Using a diagnostic testing framework, accuracy is analogous to positive predictive value (or precision in an information retrieval framework). Our gold standards were not necessarily complete (they did not contain all medications, all laboratory tests, all diseases and all associations or indications), so it was not possible to calculate sensitivity and specificity, or to carry out a receiver operating characteristic (ROC) analysis.

5. Results

Data were successfully acquired for 100,000 Brigham and Women's Hospital patients. The dataset included 272,749 coded problems, 442,658 coded medications and 11,801,068 coded laboratory results from the EHR system. There were 1756 unique coded problems, 2128 unique medications and 1341 unique coded laboratory results. The total size of the dataset was 762 megabytes (laboratory test results predominated). We ran our programs on the dataset, which took approximately nine minutes to complete (reading the data into efficient in-memory structures predominated – the actual analysis step was very short).

5.1. Medication-problem associations

A total of 10,735 medication-problem associations with support of at least 5 and confidence of at least 10% were identified. We characterized all of these pairs with the five statistics described

Table 1
Top 50 medication-problem associations under chi square.

Medication	Problem	Support	Confidence	Chi square	Interest	Conviction
Cyclosporine micro (Neoral)	Cardiac transplant	72	47.37%	15974.05	222.76	1.90
Ritonavir	HIV/AIDS[b]	108	87.10%	13584.49	126.62	7.70
Tenofovir/emtricitabine[a]	HIV/AIDS[b]	117	74.05%	12484.95	107.66	3.83
Multivitamin (vitamins A, D, E, K)	Cystic fibrosis	13	76.47%	12206.84	939.93	4.25
Atazanavir	HIV/AIDS[b]	91	87.50%	11495.76	127.21	7.94
Efavirenz/emtricitabine/tenofovir[a]	HIV/AIDS[b]	77	95.06%	10576.62	138.20	20.11
Efavirenz/emtricitabine/tenofovir[a]	HIV positive	73	90.12%	10525.03	145.06	10.06
Ritonavir	HIV positive	90	72.58%	10423.49	116.82	3.62
Cyclosporine micro (Neoral)	Stress test	63	41.45%	10390.04	166.04	1.70
Tenofovir/emtricitabine[a]	HIV positive	101	63.92%	10284.74	102.89	2.75
Atazanavir	HIV positive	79	75.96%	9579.52	122.27	4.13
Dornase alfa	Cystic fibrosis	11	68.75%	9283.65	845.03	3.20
Hydroxyurea (non-oncology dose)	Sickle cell anemia	13	44.83%	8502.43	655.23	1.81
Pancrelipase 20,000 units	Cystic fibrosis	16	41.03%	8048.56	504.26	1.69
Cyclosporine micro (Neoral)	Cardiac catheterization	65	42.76%	6597.73	102.79	1.74
Clozapine	Schizophrenia[b]	39	57.35%	6352.86	164.11	2.34
Hydroxychloroquine	Systemic lupus	204	23.26%	5863.47	30.02	1.29
Cyanocobalamin	B12 deficiency	186	18.62%	5815.46	32.48	1.22
Allopurinol	Gout	495	45.50%	5513.09	24.17	1.80
Tiotropium	COPD	223	41.22%	5430.53	25.68	1.67
Abacavir/lamivudine	HIV/AIDS[b]	40	93.02%	5371.03	135.24	14.23
Tiotropium	COPD[b]	224	41.40%	5255.21	24.80	1.68
Colchicine	Gout	244	42.00%	5117.07	22.31	1.69
Clozapine	Schizophrenia	31	45.59%	5088.16	165.47	1.83
Latanoprost	Glaucoma	218	57.07%	5044.50	24.40	2.27
Cabergoline	Prolactinoma	20	24.10%	4710.94	236.94	1.32
Pentosan polysulfate	Interstitial cystitis	13	44.83%	4617.13	356.52	1.81
Efavirenz	HIV/AIDS[b]	39	81.25%	4564.53	118.12	5.30
Abacavir/lamivudine	HIV positive	35	81.40%	4547.55	131.01	5.34
Methotrexate (non-oncology dose)	Rheumatoid arthritis	219	38.76%	4405.36	21.50	1.60
Carbidopa/levodopa[a]	Parkinson's[b]	47	33.57%	4377.79	94.56	1.50
Mesalamine	Crohns disease	54	45.38%	4373.43	82.35	1.82
Ritonavir	AIDS	20	16.13%	4228.06	212.75	1.19
Griseofulvin	Tinea capitis	5	21.74%	4191.79	839.78	1.28
Glatiramer	Multiple sclerosis	124	34.25%	4166.33	35.02	1.51
Efavirenz	HIV positive	35	72.92%	4066.94	117.37	3.67
Desmopressin nasal	von Willebrand's	7	63.64%	3876.90	555.09	2.75
Pancrelipase 16,000 units	Cystic fibrosis	5	62.50%	3834.74	768.21	2.66
Nephrocaps	End stage renal disease	39	22.16%	3742.90	97.43	1.28
Lopinavir/Ritonavir	HIV/AIDS[b]	40	63.49%	3641.93	92.31	2.72
Enofovir	HIV/AIDS[b]	32	78.05%	3594.71	113.47	4.52
Carbidopa/levodopa[a]	Parkinson's disease[b]	40	28.57%	3555.36	90.36	1.40
Carbidopa/levodopa[a]	Parkinson's disease[b]	40	28.57%	3555.36	90.36	1.40
Tocopherol-dl-alpha	Cardiac transplant	74	10.44%	3539.56	49.08	1.11
Azathioprine	Cardiac transplant	51	14.78%	3473.72	69.52	1.17
Pyridostigmine	Myasthenia gravis	10	25.64%	3451.86	346.68	1.34
Lopinavir/Ritonavir	HIV positive	37	58.73%	3449.47	94.53	2.41
Calcipotriene	Psoriasis	130	38.92%	3303.25	26.85	1.61
Furosemide	Congestive heart failure	351	11.51%	3290.76	10.66	1.12
Tenofovir	HIV positive	29	70.73%	3266.62	113.85	3.40

[a] Combination product.
[b] Problem classes.

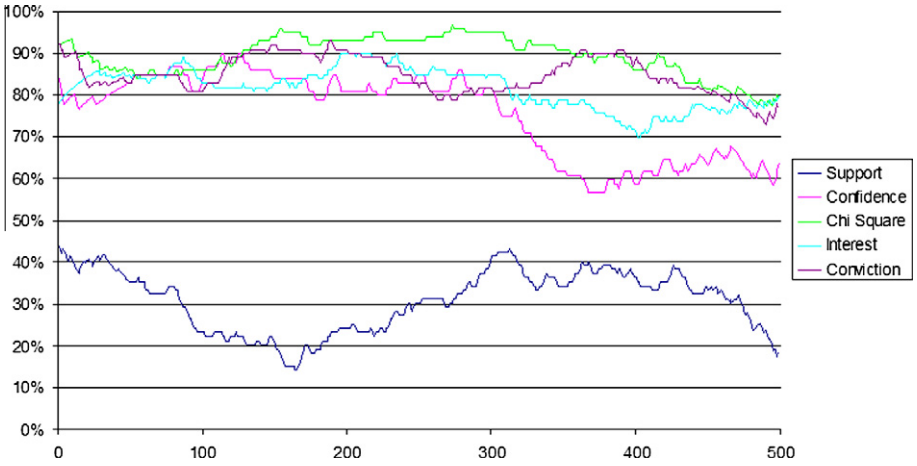

Fig. 3. Centered moving average accuracy of the top 500 medication-problem associations according to five statistics.

Table 2
Top 50 laboratory-problem associations under interest.

Laboratory Result	Problem	Support	Confidence	Chi Square	Interest	Conviction
Bethesda inhibitor assay	Hemophilia	7	25.00%	5906.03	845.03	1.33
vWF multimers	von Willebrand's disease	8	53.33%	3711.07	465.22	2.14
Fetal hemoglobin	Sickle cell anemia	18	25.35%	6647.50	370.57	1.34
Cotinine	Lung Transplant	9	18.75%	2452.46	274.06	1.23
Cotinine	Cystic Fibrosis	10	20.83%	2545.07	256.07	1.26
Cotinine	Pulmonary Fibrosis	9	27.27%	1997.01	223.48	1.37
Vitamin K	Cystic Fibrosis	8	17.39%	1696.96	213.76	1.21
Cyclosporine level	Cardiac Transplant	101	42.98%	20344.05	202.12	1.75
Tobramycin level	Cystic Fibrosis	10	16.13%	1966.38	198.25	1.19
Cotinine	Pulmonary Fibrosis	11	22.92%	2048.01	187.78	1.30
HHV6 type	Graft vs. host disease	5	10.64%	890.30	179.79	1.12
Voriconazole level	Bone marrow transplant	5	26.32%	870.16	175.71	1.36
HHV6 PCR	Graft vs. host disease	5	10.20%	853.58	172.46	1.11
Respiratory syncytial virus	Lung transplant	14	11.29%	2289.29	165.03	1.13
Cyclosporine level	Stress test	91	38.72%	14031.15	155.13	1.63
HEP C SUPPLEMENTAL	Pulmonary fibrosis	9	18.37%	1339.46	150.51	1.22
Acetylcholine receptor antibodies	Myasthenia gravis	7	11.11%	1039.62	150.23	1.12
Plasma hemoglobin	Cytomegalovirus	5	11.90%	739.84	149.73	1.13
Bone marrow aspirate	Acute myeloblastic leukemia	18	16.51%	2607.68	146.41	1.20
Vitamin A	Cystic fibrosis	10	11.63%	1412.69	142.92	1.13
Vitamin E	Cystic fibrosis	9	11.54%	1261.32	141.82	1.13
RPR titer	Syphilis	27	17.20%	3709.60	138.82	1.21
HHV6 PCR	Bone marrow transplant	10	20.41%	1345.92	136.26	1.25
HHV6 type	Acute myeloblastic leukemia	7	14.89%	912.19	132.05	1.17
vWF:RCo assay	von Willebrand's disease	23	15.03%	2981.98	131.13	1.18
Factor VIII:C	von Willebrand's disease	23	14.84%	2943.02	129.44	1.17
MHA-TP	Syphilis	27	15.98%	3443.14	128.96	1.19
HHV6 type	Bone marrow transplant	9	19.15%	1135.44	127.85	1.23
HHV6 PCR	Acute myeloblastic leukemia	7	14.29%	874.42	126.66	1.17
vWF antigen	von Willebrand's disease	23	14.29%	2831.95	124.61	1.17
Plasma hgb	Cardiac transplant	11	26.19%	1336.82	123.17	1.35
Coccidioidomycosis	Pulmonary fibrosis	10	14.71%	1188.06	120.50	1.17
HBsAg neutralization assay	Hepatitis B	6	42.86%	592.77	100.34	1.74
Adenovirus PCR	Bone marrow transplant	8	14.81%	777.34	98.92	1.17
BK virus PCR	Kidney transplant	5	12.82%	465.92	94.98	1.15
Cyclosporine level	Cardiac catheterization	92	39.15%	8546.73	94.10	1.64
Rapamycin level	Bone marrow transplant	34	13.99%	3127.36	93.42	1.16
Blasts	Acute myeloblastic leukemia	21	10.24%	1874.69	90.82	1.11
BK viral load	Kidney transplant	17	12.23%	1512.44	90.61	1.14
Epinephrine-induced plt agg (100 μm)	von Willebrand's disease	5	10.20%	436.00	89.01	1.11
Ristocetin-induced plt agglut	von Willebrand's disease	5	10.20%	436.00	89.01	1.11
Collagen-induced plt agg	von Willebrand's disease	5	10.20%	436.00	89.01	1.11
Epinephrine-induced plt agg	von Willebrand's disease	5	10.20%	436.00	89.01	1.11
Arachiodonate-induced plt agg	von Willebrand's disease	5	10.20%	436.00	89.01	1.11
FMC-7	HIV positive	246	54.67%	21467.36	87.99	2.19
Lymphogranuloma venereum ab	HIV/AIDS[a]	9	60.00%	772.69	87.23	2.48
FMC-7	HIV/AIDS[a]	264	58.67%	22329.93	85.29	2.40
CD4	HIV positive	254	50.50%	20456.20	81.28	2.01
Mycophenolic acid	Cardiac transplant	5	17.24%	396.53	81.08	1.21
CD4	HIV/AIDS[^]	273	54.27%	21342.46	78.90	2.17

[a] Problem classes.

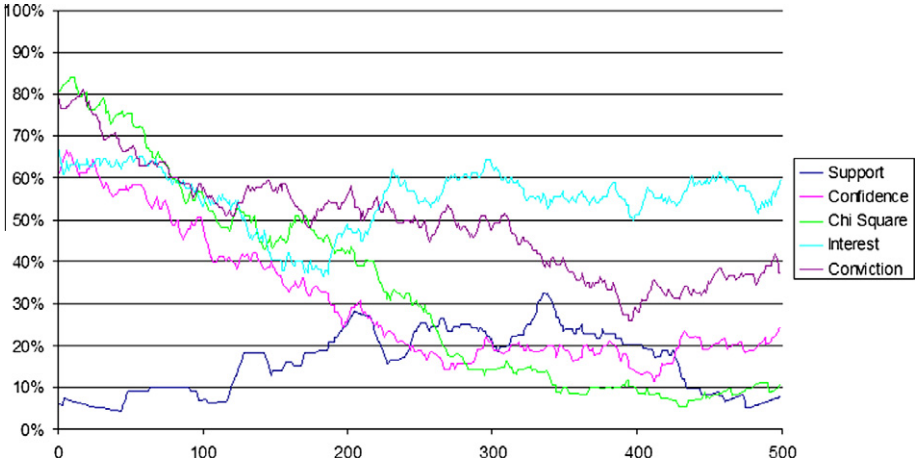

Fig. 4. Centered moving average accuracy of the top 500 laboratory-problem associations according to five statistics.

in Section 3: support, confidence, chi square, interest and conviction.

Table 1 shows the top 50 medication-problem associations based on the chi square statistic. A review of the table suggests that all 50 associations are clinically valid when compared to the gold standard, and that many of them are also very specific (for example, a variety of anti-retroviral agents are associated with HIV and/or AIDS – these agents are used only to treat HIV and AIDS). Several rows bear special mention. First, some of the problems (marked with ^) are actually problem classes (described in Section 3 and Appendix 1). In a few cases, this causes duplicate associations: for example, ritonavir is associated both with the HIV/AIDS class and the problem "HIV positive". The HIV/AIDS class association has a higher confidence (87.10%) than the HIV positive problem association (72.58%). This is because some patients have only AIDS and not HIV on their problem list, so when the two are combined the confidence increases. It should be noted that ritonavir is used only to treat HIV (with our without AIDS), so the 12.90% of ritonavir-using patients with neither HIV nor AIDS on their problem list represents an omission (accidental or intentional) from those patients' problem lists.

The pancrelipase, methotrexate and hydroxyurea associations also merit special mention. Although we did not explicitly consider dosing in our analysis, these drugs have doses imbedded in them in the order entry system. This is designed, in the case of methotrexate and hydroxyurea, to enable indication based dosing: these drugs both have oncology uses as well as non-oncology uses (rheumatoid arthritis for methotrexate and sickle cell disease for hydroxyurea) with widely differing doses. Because our system captures the indication and dose range with the order, we can pick out associations between the non-oncology uses and specific problems (we did not find specific associations for these drugs in the domain of oncology problems likely because their use in oncology is so broad).

Fig. 3 shows the results of our gold standard evaluation. We compared the top 500 medication-problem associations according to each of the five statistics to the gold standard (the Lexi-Comp drug database). Fig. 3 shows how the accuracy of the associations decays as a function of each statistic. Chi square appeared to have the best performance, consistently maintaining accuracy throughout the top 500. Support had the worst accuracy, starting strong but quickly dropping to the 30%-40% accuracy range. Of the top 500 associations, according to the chi square statistic, 89.2% were also found in the gold standard suggesting a high level of accuracy.

We conducted an analysis of the 10.8% associations that were adjudged incorrect. Although not seen in the top 50 medication-problem associations, we found that many of the apparent associations appeared to be transitive. For example, there was an association between insulin lispro 75%/insulin lispro protamine 25% mix (Humalog Mix 75/25) and hypertension. Indeed 60.9% of patients on this insulin preparation also had hypertension on their problem list, and $\chi^2 = 33.20$ for the association ($p < 0.0001$). Although this association is strictly true (indeed, clinically, most diabetic patients on insulin do have hypertension), insulin is not used to treat hypertension. The association is transitive: insulin lispro → diabetes → hypertension.

To control for these transitive associations, we used the novel iterative transitive reduction technique described in Section 3. Our method begins with calculating problem-problem associations to locate statistical comorbidities (we found 17,951 comorbidity rules with support 5 and confidence 10). Then, when we locate a potential association, such as the insulin lispro → hypertension association, we find comorbidities of hypertension (54.52% of diabetic patients in our sample have hypertension) and repeat our analyses holding out these comorbidities one-by-one. When we re-test the insulin lispro → hypertension association excluding all diabetic patients, the support drops to 1 and χ^2 falls from 33.20 to 0.13, strongly suggesting that insulin lispro → hypertension is transitively mediated by diabetes. When other comorbid conditions are used in the hold-out criteria, the chi square statistic changes very little and remains statistically significant.

5.2. Laboratory-problem associations

As mentioned in the methods, laboratory-problem associations were generated in three different ways: by test, by test with flag and by test with qualitative result. Using a support threshold of 5 and a confidence threshold of 10%, there were 5361 associations with the "by test" method, 8383 with the "by test with flag" method and 5795 with the "by test with qualitative result" method. Each of these methods had its own unique advantages. For example, the mere presence of an HIV screening test means little, but a positive result indicates a high likelihood that the patient has HIV; so in this case, the "test with flag" method would yield the best results. By contrast, the mere presence of a CD4/CD8 ratio test, regardless of the result, strongly suggests HIV because the test is ordered almost exclusively in this population so the "by test" method may work best. However, for a test with qualitative results (such as a blood smear), there are no flags, so the result itself must be used, making the "by test with qualitative result" superior.

Table 2 shows the top 50 associations using the "by test" method according to the interest statistic. The interest statistic is presented here because it had the highest accuracy (55.6% across the top 500). We focused the analysis on the "by test" method because our gold standard provided clear indications for each test, but interpretation of the test results (either by reference range driven flags or qualitative results) was much more subjective. Like Table 1, the results in Table 2 appear generally accurate based on the gold standard. The table contains a number of drug levels paired with associated conditions, some viral and bacterial antibodies and PCR tests that are highly specific for their associated problems, a number of transplant-related tests and associated transplants as well as a host of tests related to von Willebrand's disease and a substantial number of HIV-related tests.

Fig. 4 shows the results of our gold standard analysis. As mentioned in Section 3, unlike medications, where all medications were listed in our gold standard, not all laboratory tests were listed in our laboratory gold standard. As a result, each association identified by our techniques was coded "indicated", "not indicated" or "not found". The "not found" results were excluded from our analysis. The overall accuracy of the laboratory-problem associations is not as strong as the medication-problem associations and the statistics decay more quickly. However, most of the statistics start out with high accuracy and over the full run of 500, have about 50% accuracy.

6. Discussion

Overall, the techniques appear to have worked well and achieved reasonable accuracy. We were able to analyze a large amount of data in a reasonable period of time. We found that the chi square statistic had the best general performance for medications, while the interest statistic was best for laboratory results. The support statistic had the worst performance in both cases. This suggests that there is no clear "best" statistic – instead, statistics should be chosen based on individual data sets and applications – this finding has reported elsewhere in the literature [29]. Indeed, picking the optimal statistic is both an art and a science – some

statistics may be heavily biased towards frequently occurring patterns (e.g. support), while others may favor infrequent but strong associations (e.g. interest), and still others try to balance these tradeoffs. Likewise, picking the optimal cut-point for these statistics should also be done with careful reference to both the data and application. For some applications, the cost of a false positive (incorrectly inferring a problem the patient does not have) may be very high (e.g. automated initiation of treatment protocol), while for others, the cost of a false negative (failing to infer a problem that is present) may predominate (e.g. identifying potential patients for a research study, where representativeness is important and the researcher will confirm potential diagnoses).

A potential use of the medication-problem associations and laboratory-problem associations identified by these techniques is identifying and rectifying gaps in problems lists. The fact that the associations were nearly 90% accurate for medications and 50% accurate for problems suggests that, with some appropriate manual review, one could consider implementing them as rules in a clinical information system. In both cases, the results appeared to have reasonable positive predictive value, which would be important for any clinical decision support system.

6.1. Comparison to other techniques

There are alternatives to using data mining to determine relationships between medications, laboratory results and problems. One alternative is a knowledge-based technique, where human experts determine associations between medications, laboratory results and problem. The techniques used in this study have some advantages over alternative knowledge-based techniques that may be manually intensive and costly.

First, our techniques offer advantages in terms of speed and time. It took about nine minutes to process the entire data set and generate thousands of medication-problem and laboratory-problem associations. Having experts do the same thing would have been much more time-consuming. Our automatically generated associations may require manual review; however, such review is likely to be more efficient than creating rules from scratch.

Second, our technique has advantages in terms of currency: an expert-curated knowledge base must be constantly updated to account for new clinical knowledge and new clinical entities (such as novel drugs); this knowledge management task is very large and perfect currency may be nearly impossible. Our techniques, by contrast, can be repeated as often as is desired, and the incremental cost is negligible.

Third, our techniques may better reflect current practice patterns. For example, some drug knowledge bases (such as the FDA's SPL project [36]) reflect only approved uses of medications, but these techniques can infer both on- and off-label uses.

Fourth, these techniques include inherent metrics. For example, an expert might state that ''metformin is used to treat diabetes'' but assigning a certainty to this statement is difficult. Our techniques indicate that this association held only 70.6% of the time. Review of the 29.4% of patients on metformin without diabetes also indicates that a greater-than-expected proportion of them have polycystic ovarian syndrome or breast cancer (alternative uses of metformin).

The final advantage of our technique relates to terminologies. Because our techniques operate directly on EHR data, the associations we find are automatically coded using the same terminologies as the EHR. However, implementing an expert statement like ''metformin is associated with diabetes'' requires manual mapping of the metformin and diabetes concepts. This mapping is also error prone. In addition to plain metformin, a variety of metformin-containing products are also available (e.g., combined with glyburide or rosiglitazone, or an extended release formulation). All of these products are automatically flagged as related to diabetes by our techniques, but a knowledge engineer would have to know about these products and manually associate them.

Though these advantages are important, there are also some disadvantages of these techniques when compared with knowledge-based techniques. First, these techniques work best for frequently occurring combinations. There are likely many medications, problems and laboratory tests which are used so infrequently that they could not be identified by our methods despite being strongly related. Knowledge-based techniques, given sufficient resources, would be able to identify such relationships (e.g., through substantial literature review).

Second, some of the relationships found in our analysis are only indirect. While it is, for example, strictly true that insulin is strongly associated with hypertension, the association is not direct, and a knowledge base that posited that ''insulin is used to treat hypertension'' would be incorrect – an error unlikely to be made with a knowledge-based mechanism. This disadvantage may, however, also be an advantage in certain settings: if one were attempting to locate hypertensive patients, it might be reasonable to screen insulin users despite the lack of a direct relationship.

The final disadvantage is the dual of the third advantage: the techniques reflect current practice patterns. To the extent that these practice patterns may be less-than-ideal, or at least not entirely evidence-based, any application of these techniques that tends to perpetuate these patterns may have undesirable results. Knowledge-based techniques, particularly those grounded in evidence, are less likely to perpetuate sub-optimal practice patterns and may, in fact, be useful for correcting such sub-optimal patterns. However, at the same time, these techniques may be powerful tools for identifying and characterizing practice patterns (positive or negative), so that sub-optimal practice might be remediated.

Given this balance of advantages and disadvantages, it is also reasonable to imagine that association rule mining might be used in conjunction with knowledge-based techniques to create maximally effective decision support systems. For example, experts might be presented with automatically inferred association rules, and could validate (or reject them), or simply use them as another input to their knowledge base development process. Alternatively, expert-generated content could be retrospectively validated against association rules and other automatically derived measured, allowing the content to be characterized, and firing rates and accuracy to be predicted before the content goes live. Indeed, all of these approaches could be combined, iteratively, to create a more data-driven, efficient and measurable expert knowledge base development process.

7. Limitations

The techniques used and results have some limitations. First and foremost, this was a single-site study, and our site has fairly advanced clinical systems with good uptake. In a setting with less automation, lower utilization of clinical systems, or less availability of structured data, the techniques might not be as successful. We do, however, believe that the methods are highly generalizable and that similar analyses could be carried out at other sites with similar results.

Second, we limited our analysis to three data types (problems, medications and laboratory results), and only to structured information. There may be additional information available only through other data types that were not included (such as unstructured free text information, or procedure histories) and could be used to validate or complement findings.

Finally, our evaluation compared results to a fixed gold standard. This allows us to measure the accuracy of our techniques but only enables us to speculate about their utility. Since the ultimate goal of these techniques is to identify and help remediate potential gaps in clinical problem lists, an experimental evaluation with an effector arm, such as a system which alerts physicians to probable gaps and invites them to correct them, would allow for more definitive assessment of the techniques' practical utility.

8. Next steps

One obvious next step to extend this work is the addition of further structured data types. For example, procedures may be strong predictors of problems (e.g., CABG for CAD), as might certain visit types or providers (e.g., a patient who visits a mesothelioma clinic likely has mesothelioma, and a patient who visits a urologist who does only TURP likely has BPH). Reliable coded data for these data elements was not readily available to us; however, we hope to acquire and analyze such data in the future.

We also plan to extend our work to consider non-structured data, such as progress notes, radiology reports and operative notes. We believe that these data sources may contain rich predictive information, which is not always available in structured form. For example, the ejection fraction from an echocardiography report may powerfully indicate congestive heart failure, while the indication listed in an operative note might allude to a condition that the patient was treated for but might not be documented on the problem list. We have conducted a small feasibility test of these methods and the results appear intriguing. For example, our association rule mining strategy applied to outpatient notes found that the word "diabetes" in a patient note was not very strongly predictive of having diabetes, largely because of phrases like "will screen for diabetes", "no diabetes" or "family history of diabetes". However, the words "strips", "juice" and "Joslin" (a local diabetes center) were strongly associated with having diabetes. Although these associations make sense in retrospect, we might not have thought of them prospectively.

Further, there may be some value in attempting to locate larger association rules (i.e., rules with more than one antecedent or consequent). These associations may improve the specificity of rules in ways that are impossible with a single data element. For example, as discussed above, the drug metformin was associated with diabetes in our dataset and this is its primary indication; however, it is also used in the treatment of polycystic ovarian syndrome, so the association between metformin and diabetes is likely to have some false positives. However, adding laboratory information (e.g. the patient's last HbA1c value) to the antecedent set of the association might result in higher confidence (and accuracy under gold standard review). This extension would be quite powerful; however, it is challenging for two reasons: first, it would be considerably harder to find a gold standard – in our evaluation, we used reference sources for medication and laboratory result indications – a similar evaluation of implications involving combinations of multiple drugs or lab results with a single or multiple problems would require a more sophisticated reference source to sue as the gold standard (we are not aware that one exists). Further, when the Apriori algorithm is used to generate rules with large antecedent and consequent sets, the results can be "noisy", containing various trivial combinations and supersets of meaningful rules.

As a further extension of the methods, there likewise may be value in locating unknown associations, such as unexpected problem-medication linkages that could be signs of adverse drug events, or unexpected laboratory-problem associations which may be signs of potential new indications for a test. These are currently counted as "false positives" in our analysis; however, some of them may represent potentially interesting new hypotheses for more detailed investigation.

In addition to these next steps that are focused on extensions of our methods, we have also begun exploring its potential application. We are in the process of developing an intervention within our electronic health record system that will use the rules generated in this study to bring potential problem list gaps to the attention of providers and help them address these gaps. Such an intervention could have the potential to improve quality and safety as well as to enable better decision support and quality measurement.

9. Conclusion

Overall, the data mining methods described in this paper appeared to produce results with reasonable accuracy. A variety of "interesting" associations between medications and problems and laboratory results and problems were identified and described and the accuracy of these associations was verified through comparison with a gold standard. Further, if these methods can be extended and applied, they may have utility for improving problem list completeness and accuracy which may, in turn, have important benefits for patient care.

Acknowledgements

The authors are grateful to the Partners HealthCare Quality Data Warehouse and Research Patient Data Registry teams who provided the data used in the analyses. They are likewise grateful to the Partners High Performance Computing (HPC) team who provided access to the Partners HPC Cluster.

The authors also appreciate the input and advice provided by Howard Goldberg, Cheryl van Putten and Marilyn Paterno who were involved in earlier phases of this project, as well as that of David W. Bates and Gordy Schiff who provided feedback on the methods and approach. Other members of the Partners HealthCare Clinical Informatics Research and Development and Clinical Quality Analysis Groups as well as the Brigham and Women's Hospital Division of General Internal Medicine provided useful comments on the work.

This work was supported by a grant from Siemens Medical Solutions and the Partners Information Systems Research Council. Neither Partners nor Siemens had any role in the design of the study, analysis of the data, interpretation of the results, or the decision to publish.

Dr. Wright had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2010.09.009.

References

[1] Weed LL. Medical records that guide and teach. N Engl J Med 1968;278(12):652–7.
[2] Certification Commission for Healthcare Information Technology. CCHIT 2009–2010 Ambulatory EHR. 2009 [cited 2009 June 23]; Available from: <http://www.cchit.org/files/certification/09/Ambulatory/CCHITCriteriaAMBULATORY2009-2010Final.pdf/>.
[3] Hartung DM, Hunt J, Siemienczuk J, Miller H, Touchette DR. Clinical implications of an accurate problem list on heart failure treatment. J Gen Intern Med 2005;20(2):143–7.
[4] Wright A, Goldberg H, Hongsermeier T, Middleton B. A description and functional taxonomy of rule-based decision support content at a large integrated delivery network. J Am Med Inform Assoc 2007;14(4):489–96.

[5] Szeto HC, Coleman RK, Gholami P, Hoffman BB, Goldstein MK. Accuracy of computerized outpatient diagnoses in a Veterans Affairs general medicine clinic. Am J Manage Care 2002;8(1):37–43.
[6] Burton MM, Simonaitis L, Schadow G. Medication and indication linkage: a practical therapy for the problem list? Proc AMIA Symp 2008:86–90.
[7] Carpenter JD, Gorman PN. Using medication list–problem list mismatches as markers of potential error. Proc AMIA Symp 2002:106–10.
[8] Lin JH, Haug PJ. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. J Biomed Inform 2008;41(1):1–14.
[9] Poissant L, Tamblyn R, Huang A. Preliminary validation of an automated health problem list. Proc AMIA Symp 2005:1084.
[10] Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. Proc AMIA Annu Symp 2001:17–21.
[11] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(5):301–10.
[12] Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. J Biomed Inform 2006;39(6):589–99.
[13] Meystre SM, Haug PJ. Randomized controlled trial of an automated problem list with improved sensitivity. Int J Med Inform 2008;77(9):602–12.
[14] Jao C, Hier D, Galanter W. Automating the maintenance of problem list documentation using a clinical decision support system. Proc AMIA Symp 2008:989.
[15] Goethals B. Survey on frequent pattern mining. 2003 [cited 2009 June 15]; Available from: <http://www.cs.helsinki.fi/u/goethals/publications/>.
[16] Sarawagi S, Thomas S, Agrawal R. Integrating association rule mining with relational database systems: alternatives and implications. Data Mining Knowledge Discov 2000;4(2):89–125.
[17] Iskander J, Pool V, Zhou W, English-Bullard R. Data mining in the US using the vaccine adverse event reporting system. Drug Saf 2006;29(5):375–84.
[18] Carrino JA, Ohno-Machado L. Development of radiology prediction models using feature analysis. Acad Radiol 2005;12(4):415–21.
[19] Kohavi R, Mason L, Parekh R, Zheng Z. Lessons and challenges from mining retail e-commerce data. Mach Learn 2004;57(1):83–113.
[20] Cao H, Markatou M, Melton GB, Chiang MF, Hripcsak G. Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. Proc AMIA Symp 2005:106–10.
[21] Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc 2009;16(3):328–37.
[22] Mullins IM, Siadaty MS, Lyman J, et al. Data mining and clinical data repositories: insights from a 667, 000 patient data set. Comput Biol Med 2006;36(12):1351–77.
[23] Brossette SE, Sprague AP, Hardin JM, Waites KB, Jones WT, Moser SA. Association rules and data mining in hospital infection control and public health surveillance. J Am Med Inform Assoc 1998;5(4):373–81.
[24] Doddi S, Marathe A, Ravi SS, Torney DC. Discovery of association rules in medical data. Medical informatics and the Internet in medicine 2001;26(1):25–33.
[25] Chen ES, Cimino JJ. Automated discovery of patient-specific clinician information needs using clinical information system log files. Proc AMIA Symp 2003:145–9.
[26] Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inform Assoc 2008;15(1):87–98.
[27] Wright A, Sittig DF. Automated development of order sets and corollary orders by data mining in an ambulatory computerized physician order entry system. Proc AMIA Symp 2006:819–23.
[28] Agrawal R, Imielinski T. Mining association rules between sets of items in large databases. Proc 20th Int Conf Very Large Data Bases 1993:688–92.
[29] Tan P, Kumar V, Srivastava J. Selecting the right interestingness measure for association patterns. Proc Eighth ACM SIGKDD Int Conf Knowledge Discov Data Mining 2002.
[30] Brin S, Motwani R, Ullman J, Tsur S. Dynamic itemset counting and implication rules for market basket data. Proc ACM SIGMOD Int Conf Manage Data 1997:255–64.
[31] Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. Proc AMIA Annu Symp 2001:662–6.
[32] Forrey AW, McDonald CJ, DeMoor G, et al. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. Clin Chem 1996;42(1):81–90.
[33] Liu S, Ma W, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. IT Prof 2005;7(5):17–23.
[34] Narayanasamy V, Mukhopadhyay S, Palakal M, Potter DA. TransMiner: mining transitive associations among biological objects from text. J Biomed Sci 2004;11(6):864–73.
[35] Pagana KD, Pagana TJ. Mosby's diagnostic and laboratory test reference. St. Louis, Mo.: Mosby Elsevier; 2007.
[36] United States Food and Drug Administration. Structured Product Labeling Resources. 2009 [cited 2009 June 15]; Available from: <http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/default.htm/>.