On Interestingness Measures for Mining Statistically Significant and Novel Clinical Associations from EMRs

Orhan Abar
Dept. of Computer Science

Richard J. Charnigo
Department of Biostatistics

Abner Rayapati
Department of Psychiatry

Ramakanth Kavuluru*
Div. of Biomedical Informatics
Dept. of Internal Medicine
Dept. of Computer Science

University of Kentucky, Lexington, KY.
{orhan.abar, richard.charnigo, abner.rayapati, ramakanth.kavuluru}@uky.edu

ABSTRACT

Association rule mining has received significant attention from both the data mining and machine learning communities. While data mining researchers focus more on designing efficient algorithms to mine rules from large datasets, the learning community has explored applications of rule mining to classification. A major problem with rule mining algorithms is the explosion of rules even for moderate sized datasets making it very difficult for end users to identify both statistically significant and potentially novel rules that could lead to interesting new insights and hypotheses. Researchers have proposed many domain independent interestingness measures using which, one can rank the rules and potentially glean useful rules from the top ranked ones. However, these measures have not been fully explored for rule mining in clinical datasets owing to the relatively large sizes of the datasets often encountered in healthcare and also due to limited access to domain experts for review/analysis. In this paper, using an electronic medical record (EMR) dataset of diagnoses and medications from over three million patient visits to the University of Kentucky medical center and affiliated clinics, we conduct a thorough evaluation of dozens of interestingness measures proposed in data mining literature, including some new composite measures. Using cumulative relevance metrics from information retrieval, we compare these interestingness measures against human judgments obtained from a practicing psychiatrist for association rules involving the *depressive disorders* class as the consequent. Our results not only surface new interesting associations for depressive disorders but also indicate classes of interestingness measures that weight rule novelty and statistical strength in contrasting ways, offering new insights for end users in identifying interesting rules.

*corresponding author

BCB'16, October 2–5, 2016, Seattle, WA, USA.
 ISBN 978-1-4503-4225-4/16/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2975167.2985843.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*Contingency table analysis*

General Terms

Design, Measurement, Experimentation

Keywords

association rule mining, rule interestingness measures, electronic medical records

1. INTRODUCTION

Association rule mining (ARM) [1] has emerged as an important methodology to gain insights into large databases of transactions each of which contains a set of items. ARM first gained popularity for market-basket analysis where each transaction consists of a set of products purchased by a customer. Using ARM, rules of the form $E \Rightarrow Y$ are extracted which indicate that a customer that buys a set of items E "tends" to buy items in Y in the same visit. Association rules (ARs) obtained for this domain have been used to better design product placement layouts in stores that encourage so called cross-selling among customers. Similar strategies are also being employed by online stores to dynamically generate product recommendations based on prior browsing/purchasing history. In the context of biomedicine and healthcare, ARM has also been applied to EMR data for association analysis among biomedical and clinical variables [2, 19, 30, 31]. Before we proceed further, we establish some primitives for ARM starting with the notion of a clinical item set.

1.1 ARM Basics

Let $\mathcal{I}$ be union of all medications and diagnoses that can be used for patients. For our purposes, a set $E = \{i_1, \ldots, i_k\} \subseteq \mathcal{I}$ is called a clinical *item set* with k items and a patient *visit transaction* $T = (pid, vid, I)$ is defined over $\mathcal{I}$ where *vid* is the patient visit ID, *pid* is the patient ID, and $I \subseteq \mathcal{I}$ is the item set corresponding to the current visit *vid*. The set of all visit transactions in a given database is denoted as the visit database $\mathcal{V}$. A visit trans-

action (pid, vid, I) is said to *support* an item set E if $E \subseteq I$ and the *support* of E in the database $\mathcal{V}$ is defined as

$$support(E, \mathcal{V}) = |\{vid : (pid, vid, I) \in \mathcal{V}, E \subseteq I\}|.$$

An item set is deemed *frequent* if its support is greater than a given minimum support σ. Thus the set of frequent item sets with respect to σ is defined as

$$\mathcal{F}(\mathcal{V}, \sigma) = \{E : support(E, \mathcal{V}) \geq \sigma\}.$$

Next, an AR is a rule of the form $E \Rightarrow Y$ where E and Y are item sets and $E \cap Y = \emptyset$. The *confidence* of an association rule $E \Rightarrow Y$ denoted by

$$conf(E \Rightarrow Y, \mathcal{V}) = \frac{support(E \cup Y)}{support(E)},$$

models the probability $P(Y|E)$ and establishes the association of the consequent item set Y with the antecedent item set E. Like minimum support for item sets, we can establish a minimum confidence γ for ARs and define a stronger notion of frequent and confident ARs over a visit database $\mathcal{V}$ as the set

$$\mathcal{R}(\mathcal{V}, \sigma, \gamma) = \{E \Rightarrow Y : E \cup Y \in \mathcal{F}(\mathcal{V}, \sigma), conf(E \Rightarrow Y) \geq \gamma\},$$

which consists of confidence thresholded ARs obtained from frequent item sets. From a biomedical perspective, we can filter ARs $\mathcal{R}(\mathcal{V}, \sigma, \gamma)$ choosing interesting and meaningful consequents Y. For example, we can set $Y = \{$Non-Small Cell Lung Cancer (NSCLC)$\}$, that is, a consequent with just one item, NSCLC, which corresponds to patient visits that had a diagnosis code for NSCLC.

As their name indicates, ARs are essentially associations (or correlations) and do not indicate causality, although they have been known to manifest when there is a causal relationship. ARs are also used as starting points to arrive at potential causal relations [8] using additional retrospective analyses involving confounding factors (not all of which maybe recorded in a clinical database) or additional prospective experiments such as randomized control trials (which may not be feasible in all situations) [24]. We emphasize that the scope of this paper is assessing rule interestingness measures in the context of *ranking large AR sets to enable discovery of interesting associations* that can lead to novel hypotheses. Next we discuss the notions of statistical strength, novelty, and interestingness of rules generated by ARM.

1.2 Notions of Statistical Strength, Novelty, & Interestingness

Statistical significance and novelty are two important and complementary notions that make a rule desirable for further examination. Generally speaking, an AR is deemed statistically significant if its manifestation is not due to random chance. Statistical strength is a measure-specific notion that attributes a gradation or degree to the significance of the rule. Thus, we would at least want a rule to be statistically significant and also prefer for it to have high statistical strength. However, statistically significant ARs may not be meaningful or clinically relevant; even in cases when they are meaningful, they might be too obvious. For example, in our experiments, the association of antidepressants with depressive disorders is statistically significant but is very obvious to most end users. For ARM, the notion of novelty indicates the level of unexpectedness, surprise, or peculiarity associated with a rule. For example, the association between antidepressants and depressive disorders is considered not novel. For our current effort, to keep the terminology simple, novelty implicitly also includes the notion of clinical relevance or plausibility. In data mining literature [4, 26, 27, 29], "interestingness" has been used as an umbrella term to describe a combination of desirable rule properties including statistical strength and novelty and we employ the same usage for the rest of our paper. Although novelty is sometimes considered a subjective measure, in this paper we assess how various interestingness measures model novelty. Next we outline our main contributions.

1.3 Our Contributions

Prior results on applying ARM to clinical datasets [30,31] offer important insights but are based on relatively smaller datasets with a focus on rediscovering known associations already recorded in external knowledge bases. Hence they do not directly assess the novelty of the associations found. Furthermore, their evaluations consider only few interestingness measures (up to five) in their experiments and also limit the antecedent of an association to be a singleton. In our current effort

1. We use a dataset of diagnoses and medications from over 3 million patient visits to the University of Kentucky (UKY) medical center and its affiliated clinics to obtain all ARs with singleton consequents and having minimum support 100 and minimum confidence 10%. We do not limit the rule antecedents to be singletons; they can be combinations of both diagnoses and medications.

2. We rank the specific set of rules with *depressive disorders* as the consequent using over 40 different interestingness measures including most measures introduced in data mining literature [4] and a few new measures we introduce in this paper.

3. We obtain manually assigned novelty scores (1 – 5) for the set of rules in the union of top 100 rules from rankings produced by all interesting measures using the help of a practicing psychiatrist (Dr. Rayapati, a co-author of this paper). We *combine* these novelty scores and odds ratio lower bounds (from 95% confidence intervals) for these rules to compare against all interestingness measures and identify classes of measures that trade-off novelty and statistical strength in contrasting ways. We also discuss the clinical plausibility of several novel associations identified in our analysis.

The central premise for all our work is to pick specific diseases of interest as consequents and identify groups of medications and other conditions (as antecedents) that are associated with them. The associations may themselves manifest due to comorbidity situations (if antecedents are diseases). They can be indicative of treatment relations or side-effect/adverse-reaction scenarios (if the antecedents are medications). Combinations of medications and diseases as antecedents can represent more nuanced and specific scenarios with high statistical strength.

2. AR MINING FROM VISITS DATA

Here we primarily discuss the clinical dataset and methods used to extract ARs.

2.1 Clinical Dataset Used

Our dataset is extracted from all patient visits ($\approx$ 3.25 million) during the ten year period 2004-2013 to the UKY medical center and its affiliated clinics. Each visit transaction consists of medications and diagnoses recorded during a particular patient visit[1]. We also removed nearly 12,000 transactions that are very large (with 35 or more elements per visit). Although rare and in this case constituting only 0.3% of the full dataset, presence of such long transactions renders existing approaches to ARM impractical given they all rely on generating frequent item sets as an intermediate step. Thus we are still left with $\approx$ 3.25 million visits from around 572,000 unique patients. Thus, on average, each patient had about 5.66 visits during the decade. Given the ten year window of the study, we chose to treat different visits by the same patient as giving rise to different transactions. This way, the co-occurrences of medications and diagnoses are guaranteed to have the same time stamp in all our transactions.

The dataset has 11,877 unique *International Classification of Diseases, Clinical Modification, Version 9* (ICD-9-CM) codes and 1032 unique medication codes by *Cerner Multum*^{TM} *Lexicon Plus* codes which are also used by Centers for Disease Control and Prevention (CDC) for their medical care surveys. Current ARM approaches, even with the advent of "big data" approaches, do not scale well to thousands of unique items for patient visit databases with large transaction sizes especially if the minimum confidence and threshold are chosen to be small, which is critical to surface novel associations; high support and confidence rules may satisfy statistical strength requirements but tend to represent common knowledge for most end users. At lower thresholds, scalability issues mostly arise because of the combinatorial explosion of possible antecedent sets. Furthermore, considering all unique codes may not offer enough statistical strength (due to sparsity) or yield informative rules (for manual AR interpretation). For example, researchers might be more interested in knowing statistically significant and novel associations of penicillins with other conditions rather than be subjected to a deluge of weak associations involving specific penicillins such as Amoxicillin, Ampicillin, and Dicloxacillin. However, sparsity issues may be overcome by working with much larger datasets compared to the dataset used in our current effort.

Given above scenarios, we group diagnosis and medication codes using conventional approaches. For diagnoses, we use ICD-9 code classes [7] developed by the Healthcare Cost and Utilization Project (HCUP), an affiliate of the Agency for Healthcare Research and Quality (AHRQ) in the US Department of Health and Human Services. These classes group related codes resulting in 282 classes for the 11,877 codes in our dataset. For example, the HCUP class for *cancer of breast* groups 13 different ICD-9 codes covering all female breast cancer codes, male breast cancer codes, and a code for personal history of breast cancer. We rolled-up the Multum medication codes using their class hierarchy which resulted in 150 classes (e.g., Penicillins). In each transaction, we then replaced the codes with the corresponding HCUP and Multum classes resulting in a total of 432 unique items (HCUP and Multum classes) populating 3.25 million transactions.

[1]Although other variables such as procedures and labs are available, for computationally tractability we limited our current study to medications and diagnoses.

2.2 Generating Association Rules

Although there are several efficient implementations that extract frequent item sets [6, 33], including those that work on big datasets using MapReduce [17], for our purposes the Linear-time Closed item set Miner (LCM Ver. 3) by Uno et al. [28] that exploits a clever combination of bitmaps, prefix trees, and array lists worked best. We used a minimum support $\sigma = 100$ and confidence $\gamma = 10\%$ for singleton consequent rule generation. That is, in each AR, we require that the antecedent items and consequent co-occur at least 100 times in over 3 million transactions and at least 10% of the transactions that contain the antecedent set also include the consequent. This is in line with other efforts [30, 31] on applying ARM to clinical datasets. LCM generated nearly 22 million rules for our dataset.

At this point, to evaluate interestingness measures for both statistical strength and novelty, we needed to pick a narrow focus. According to the National Comorbidity Survey Replication (2001–2003), 68% of adults with mental disorders have medical conditions and 29% with medical conditions have mental disorders [12]. A February 2011 Robert Wood Johnson Foundation (RWJF) research synthesis report [3] presents evidence that this subgroup of people with mental and medical disorder comorbidities are at significant risk for poor quality of care and high costs. Depressive disorders are one of the most common mental disorders especially among adults and hence we picked the corresponding HCUP class for our focused study. The *depressive disorders* HCUP class has sixteen ICD-9 codes, which represent all variants of depression in ICD-9-CM. Our dataset has 54,923 transactions with a depressive disorder code. Post filtering all rules with depressive disorders as the consequent, we obtained 126,540 rules. Upon on manual observation, many of these rules had *antidepressants* as an element of the antecedent. Since the presence of this well known drug class that treats depression leads to uninteresting associations, we removed those rules with antidepressants as part of the antecedent, which resulted in 75,465 rules. These are the rules we ranked based on different interestingness measures.

3. ASSESSING INTERESTINGNESS MEASURES FOR AR RANKING

We ranked all the 75,465 rules with depressive disorders as the consequent class using nearly three dozen probability based objective interestingness measures from a recent survey by Geng and Hamilton [4, Table IV]. This list includes popular measures such as confidence, lift, conviction, odds ratio, and information gain. Additionally, we added the χ^2-measure as it is well known for studying statistically significant associations [5,30]. We also introduced some new measures which we describe here.

3.1 Additional Interestingness Measures

To model novelty, we introduce the notion of Average Inverse Rule Frequency (AIRF) for a given AR $E \Rightarrow Y$. Recall from Section 1.1, $\mathcal{R}(\mathcal{V}, \sigma, \gamma)$ represents the set of ARs for the visit databases $\mathcal{V}$ satisfying minimum support σ and confidence γ. Let $\mathcal{R}^Y \subseteq \mathcal{R}(\mathcal{V}, \sigma, \gamma)$ be the set of rules with Y

as the consequent from the full set of rules, assuming the database $\mathcal{V}$, σ, and γ are fixed. We define

$$AIRF(E \Rightarrow Y) = \frac{\sum_{x \in E} \frac{|\mathcal{R}^Y|}{|\{R:\ R \in \mathcal{R}^Y \ \wedge\ x \text{ is in antecedent of } R\}|}}{|E|}.$$

Inverse rule frequency is analogous to inverse document frequency (IDF) in the TF-IDF term weighting scheme popular in information retrieval. The higher the AIRF of a rule $E \Rightarrow Y$, the fewer are the rules that contain elements of E as part of their antecedents – in this sense, rules with higher AIRF are expected to be novel/peculiar. The rationale for AIRF follows from the justification for IDF [21].

Odds ratio (OR) is a well known measure for studying associations in epidemiology[2] and more specifically, the odds ratio lower bound (ORLB) [18] of the 95% confidence interval around sample OR is used as an important measure for assessing statistical significance or lack thereof. $ORLB > 1$ indicates a statistically significant association with higher values indicating stronger associations. Our new measures of interestingness for a rule $E \Rightarrow Y$ include its $AIRF$, $ORLB$,

$$\frac{ORLB}{\log_2(|E|+|Y|)},\ \textit{and}\ \frac{AIRF \cdot ORLB}{\log_2(|E|+|Y|)}, \tag{1}$$

where $\log_2(|E|+|Y|)$ indicates the length of the rule. (Note $|Y| = 1$ for our purposes and the expression equals 1 for singleton associations where additionally $|E| = 1$). Given ORLB indicates statistical strength and AIRF models novelty, we combined both in the product measure. Although we support longer rules with $|E| > 1$, very long rules are not interesting as they capture highly specific scenarios that are not amenable to reasonable interpretation and typically have low support as noted in prior efforts [5]. At the same time we do not want to severely discount long rules. So to prefer smaller rules and dampen the effect of the length on overall interestingness score, we use $\log_2(|E| + |Y|)$ in the denominator of the two measures in equation 1.

3.2 Domain Expert Novelty Assessments

We used ORLB introduced in Section 3.1 as a proxy for statistical strength in our final assessment of all interestingness measures given it is routinely considered in biostatistics. The rationale for using ORLB over OR is that ORLB balances assurance that the result is not due to chance with the strength of the estimated effect, considering the variance of the estimator. However, we do not have a similar measure for novelty. Given it is unrealistic to have domain expert assessments on 75,000 rules we combined the top 100 rules from each of the rankings produced by all interesting measures discussed in this section. That is, given $\mathcal{M}$ is the set of all interestingness measures, human annotations are assigned to the set of rules

$$\bigcup_{m \in \mathcal{M}} Rank_m^{100}(\mathcal{R}^Y), \tag{2}$$

where $Rank_m^k$ indicates a function that returns the top k rules (without any limitations on rule length) obtained by ranking using measure m. In addition to this, all singleton antecedents which had an $ORLB > 1$ were also presented to the domain expert. We did this because singleton associations ($|E| = 1$) are easier to interpret, relatively very few compared to longer rules, and $ORLB > 1$ already indicates statistically significant association.

Novelty ratings were assigned on a scale of 1 to 5 (with 5 indicating most novelty) by a practicing psychiatrist from the university's department of psychiatry. As we indicated earlier in Section 1.2, the notion of novelty (regardless of the degree) for our purposes includes plausibility. So a rating of 1 for a rule indicates it is a well-known association whose underlying mechanism is also reasonably understood. On the other hand a rating of 5 means it is a highly novel rule that is also clinically plausible although the details of the mechanism may not be as clear as for a rule with rating 1. This is to be expected given high novelty usually also implies that pertinent broad knowledge is lacking (see Section 4 for literature search based evidence for this). The assessments are informed by the physician's general medical knowledge and experiences as a practicing psychiatrist. Besides the actual rules, no other information was provided to the physician, who was requested to provide additional qualitative feedback on associations that were deemed highly novel. We chose the top 100 rule set union from all measures in the interest of domain expert time needed for novelty assessment. This limit has resulted in over 550 rules and we believe choosing larger thresholds could help for future efforts.

3.3 Comparison of Interestingness Measures

Next we compare interestingness measures discussed in this section across two dimensions, statistical strength and novelty, using rule ORLBs and psychiatrist assigned novelty scores as corresponding proxies, respectively. Using each interestingness measure, we rank all rules in equation 2 and any other singletons with an $ORLB > 1$ for the depressive disorders consequent. For reviewing convenience for the domain expert and subsequent analysis, we split all these rules into singleton and non-singleton antecedent rules. We ended up with a total of 231 singleton rules and 334 non-singleton rules each of which was assigned a novelty score (1–5).

The normalized discounted cumulative gain (NDCG) [9, Sections 2.2–2.3] is a popular rank quality metric in information retrieval (IR). It is typically used for search engines to measure the *gain* in terms of graded relevance of retrieved documents where relevant documents higher up in the ranking are given more weight compared with those that come later in the ranking. For interestingness measure comparison in our effort, we adapt NDCG to suit our purposes and compute normalized discounted cumulative novelty (NDCN) (from expert assigned scores) and normalized discounted cumulative ORLB (NDCO) based on the rule ranking produced according to each measure. Instead of the relevance judgment score of a retrieved document, we used a rule's novelty score (for NDCN) and ORLB (for NDCO). Besides this replacement of relevance scores with novelty and ORLB values, the exact expression used for NDCN and NDCO is identical to that of NDCG [9, Equation (2)]. We then sorted all measures based on the corresponding NDCN and NDCO values to identify best measures from the perspective of novelty and statistical strength. Like NDCG, the normalization

[2]For prospective studies, relative risk (RR) is a more intuitive measure of association strength, but OR is a symmetric measure that is typically used for retrospective studies and approximates RR for rare outcomes [22, Chapter 13.3]. The advantage of OR over RR is that OR can be validly estimated whether random samples are drawn from the population as a whole, from exposure/risk factor strata, or from outcome strata.

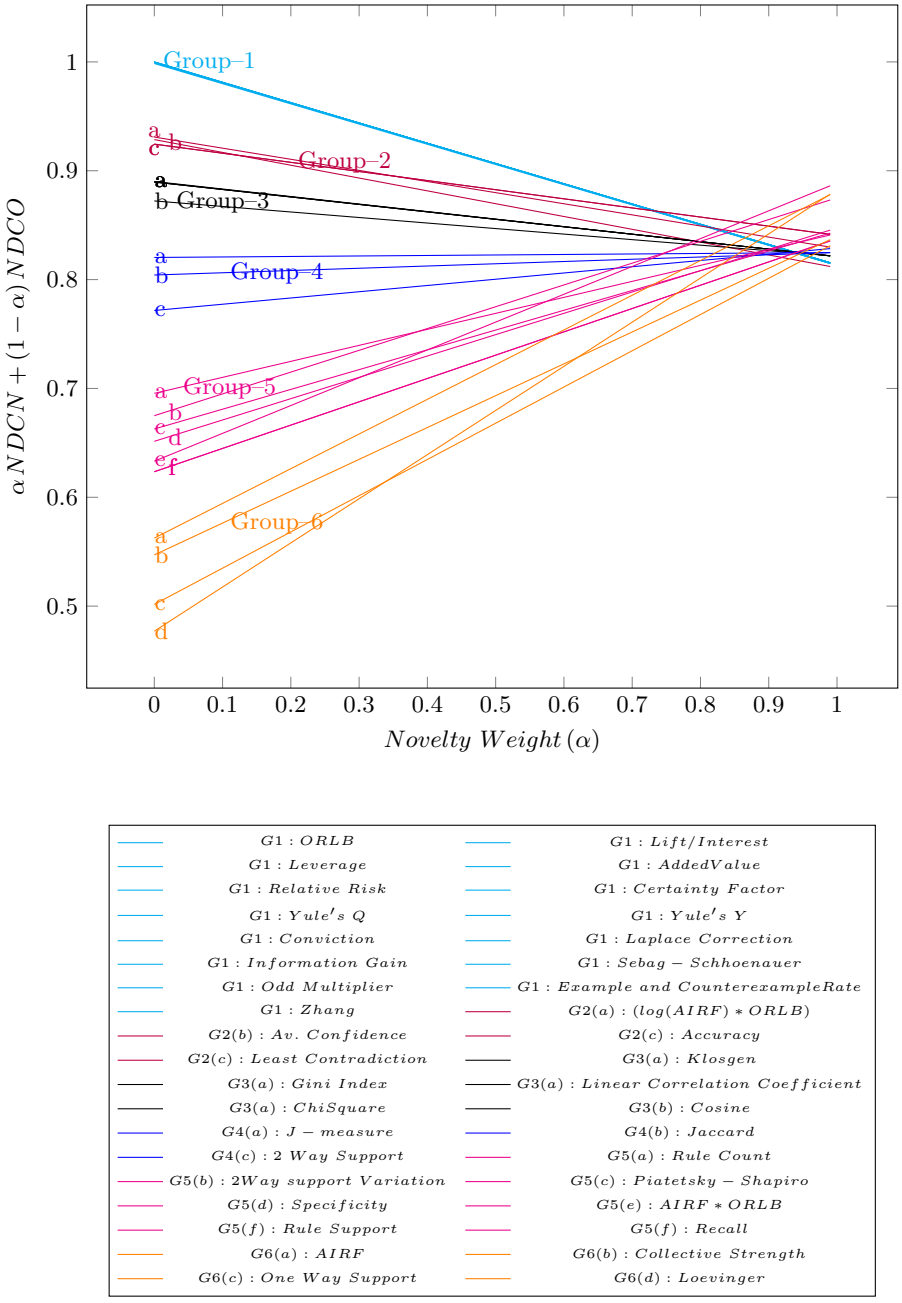

Figure 1: Interestingness Measure Profiles with Novelty-Statistical Strength Trade-Offs

aspect of our formulations implies both NDCN and NDCO take values in $[0, 1]$, where a value to closer to 1 indicates higher rank quality.

Instead of a single measure, we found classes of measures that scored similarly based on NDCN and NDCO values. Specifically, for singleton rules, $AIRF \cdot ORLB$ gave the highest NDCN value of 0.88. Measures such as $AIRF$, Loevinger, and 2-way support variation [4, Table IV] had NDCN values in $[0.87, 0.88]$. The lowest values for NDCN resulted from measures such as relative risk, Yule's Q, information gain, lift/interest, and conviction [4, Table IV] all of which had NDCN value around 0.81. On the other hand, for NDCO, these measures gave the maximum values of around 0.99. Similarly, Loevinger, which is among the top scorers for NDCN, generated the lowest NDCO score of 0.47. This demonstrates the clear trade-off between statistical strength and novelty in terms of what several interestingness measures are trying to capture.

To further compare the measures where different levels of importance are given to novelty (vs statistical strength), we plotted a combination metric

$$\alpha \cdot NDCN + (1-\alpha) \cdot NDCO \quad \in [0, 1]$$

for all measures for $\alpha = 0, 0.01, 0.02, \ldots, 0.99, 1$. The results of this plot are shown in Figure 1. For convenience, we divided the measures into high level groups and appropriate

subgroups with memberships as indicated in the legend of the figure. First we consider the six (Group1–6) different high level groups of measures with their corresponding performance profiles as α is varied. These groups were identified based on how they cluster together when statistical significance is solely considered (that is, when $\alpha = 0$). Group-1 has fifteen measures and is heavily biased toward maximizing statistical strength but also represents the top set of measures even when assigning equal importance to novelty and strength ($\alpha = 0.5$). Group-4's performance is relatively stable but does not generate superior overall performance. Group-5 archives novelty values that are higher than those of groups 1, 3, and 4. If we look at the measures from a novelty perspective, they break down into two distinct groups as can be observed when $\alpha = 1$ on the right most end of the plot in Figure 1. The first group achieves higher NDCN values and has four measures: $AIRF \cdot ORLB$ (G5(e)), $AIRF$ (G6(a)), Loevinger (G6(d)), and two way support variation (G5(b)). The rest of the measures can be clustered into the second group. Depending upon a particular user's specific preferences toward strength and novelty, he/she can choose an appropriate measure based on variations noticed in the figure. When $\alpha = 0$, measures in Group-1 are recommended; but to maximize novelty ($\alpha = 1$), measure $AIRF \cdot ORLB$ appears superior.

For longer rules, the highest NDCN of 0.933 was achieved by $ORLB/\log_2(|E| + |Y|)$, where $|E| + |Y|$ represents the length of the rule. However, several other measures such as relative risk, Yule's Q, information gain, lift/interest, and conviction all had NDCN very close to 0.93. For NDCO, the highest value of 1 was achieved by Yule's Q and Yule's Y (besides ORLB). Other measures that scored well for NDCN also scored close to the maximum value for NDCO. Hence for longer rules, the trade-off effect that led to different groups of measures that lean toward either novelty or statistical strength does not seem to exist.

4. QUANTITATIVE & QUALITATIVE ANALYSIS OF NOVEL RULES

We took two different approaches to analyze rules that were judged novel by the domain expert. We first manually mapped the medications and disease classes to Medical Subject Headings (MeSH terms), which are used to categorize biomedical articles by the US National Library of Medicine (NLM). Our visit item to MeSH mapping was done based on simple look-ups of the item names in the MeSH browser (`https://www.nlm.nih.gov/mesh/MBrowser.html`) and with the assistance of NLM's Unified Medical Language System (UMLS) to identify synonymous names. Since some HCUP and medication classes have multiple related items, some of them translated to multiple MeSH terms. MeSH terms are typically used to search biomedical articles using NLM's PubMed web application. For a given singleton rule $\{e\} \Rightarrow \{y\}$, we searched PubMed with the Boolean query

$$\left(\bigvee_{t1 \in MeSH(e)} t1\right) \bigwedge \left(\bigvee_{t2 \in MeSH(y)} t2\right)$$

for items e and y where $MeSH(x)$ denotes the MeSH term set for item x. For those singleton rules with expert assigned novelty scores ≤ 3 (total: 170), we retrieved an average of 1168 articles per rule, but the corresponding average over rules with novelty scores ≥ 4 (total: 61) is 264 and for those rules that have the top score five (total: 17), the average is 70 articles. This clearly shows that expert assigned scores seem to be aligned with what is reported in scientific literature based on co-occurrence analysis. For example, the drug class proton pump inhibitors (PPIs) has ORLB 9.98 and pulmonary heart disease has ORLB 3.07. Both were assigned a novelty score of 4 for their association with depression. For the corresponding conjunctive queries with depression, one article was returned per query, but in both cases manual review of the articles revealed no explicit discussion of the associations. For PPIs, a similar association was found with myocardial infarction by Shah et al. [25] in a recent effort. Our findings regarding rheumatoid arthritis (ORLB: 2.37) and osteoarthritis (ORLB: 4.07) are also inline with a recent and thorough study [23] that specifically looked into the impact of 24 chronic conditions on diagnosis of major depressive disorder, which differs in some aspects from the HCUP depressive disorders class used in our effort.

Table 1: Antecedents with novelty ≥ 4 and ORLB ≥ 5

Antecedent	Novelty	ORLB
CNS stimulants	5	7.65
Antianginal agents	5	7.21
Acute posthemorrhagic anemia	5	6.64
Endometriosis	5	6.46
Somatoform disorders	4	12.33
Antacids	4	9.82
ACE inhibitors	4	8.35
Anticoagulants	4	8.27
Hormonal antineoplastics	4	8.23
Esophageal disorders	4	8.02
Muscle relaxants	4	7.28
Antiplatelet Agents	4	6.77
Leukotriene modifiers	4	6.73
Immunostimulants	4	6.52
Quinolones	4	6.24

Next, based on direct inputs from the domain expert, we comment on the clinical plausibility of some of the high scoring (novelty score 4 or 5) associations for depressive disorders. Novel associations with depression are identified for conditions such as anemia (ORLB: 6.64), asthma (ORLB: 4.83), congestive heart failure (ORLB: 4.54), coronary atherosclerosis (ORLB: 3.75), and pulmonary heart disease. All these conditions can compromise oxygen flow to the brain and can contribute to microvascular injury in white matter and contribute to atypical depression. Parkinson's disease (ORLB: 5.3) and migraine (ORLB: 3.48) affect the brain and their treatments will more than likely disrupt neurotransmitter systems implicated in depression. Behavioral disorders such as ADHD (ORLB: 8.1), oppositional defiant disorder (ORLB: 15.6), and conduct disorder (ORLB: 7.73) occur in the context of unclear biological vulnerability

and psychological constructs of low self-esteem which tend to perpetuate social chaos similar to the individual's own developmental experience. Such social stress factors (poverty, unemployment, inconsistent employment, legal consequence, substance use, divorce, psychological trauma) have also been implicated in depressive disorders. So far in this section, we have looked at 13 singleton novel antecedents with some reflection on clinical relevance. In Table 1 we show the remaining novel (score ≥ 4) associations with ORLB ≥ 5.

There were a significant number of non-singleton associations with depression where the antecedent involves the suicide and intentional self-inflicted injury HCUP class along with other conditions and medications. For instance, the combination of the suicide HCUP code with osteoarthritis had ORLB over 150 but is peculiar and could be due to the observed but not thoroughly understood link between inflammation (conditions with the "itis" suffix) biomarkers and depression. Similarly, the association of suicide and alcohol related disorders with depression is well known but when epilepsy is added as a third condition to the antecedent, the association becomes statistically much stronger but also novel given seizures (from epilepsy) are considered therapeutic for mood disorders. Given seizures are also a complication in alcohol withdrawal, epilepsy might be indicating a more complex exacerbating alcohol related disorder.

5. CONCLUDING REMARKS

With innovations in computer science, informatics, and health information technology, EMR data from healthcare facilities and claims data from private and government sponsored insurance programs have become very rich sources for mining new insights for disease prevention and treatment. ARM has shown promise in other fields and is currently being actively explored for biomedicine to generate new hypotheses and also to build interpretable predictive models. An important concern in this era of big-data is dealing with vast number of rules output by ARM methods. In this paper, we evaluate over 40 interestingness measures (including some new measures) for effective ranking of ARs across two desirable properties of statistical strength and novelty. Using domain expert assigned novelty scores and ORLB for statistical strength, we adapted information retrieval metrics to assess various interestingness measures and identified classes of measures that seem to inherently weight novelty and statistical strength in contrasting ways. End users can utilize a particular class of measures depending on their goals that might influence their preferences for novelty and statistical strength. We conducted quantitive and qualitative analyses of some of the novel associations obtained as part of this effort. To our knowledge, this is the first effort to conduct a broad scoped comparative analysis of interestingness measures for clinical ARM involving subject matter expert driven novelty assessment.

Although our effort offers a reasonable proof of concept for measure assessment, we believe additional analyses are needed based on domain expert assessments for rules with other chronic diseases as consequents. Insurance claims based datasets are at least an order of magnitude larger than our academic hospital's three million patient visits. In the future, we will work with such very large datasets which will allow us to mine fine grained ARs without collapsing diagnoses and medications to their class levels. We will also add procedures and labs to the transactions. At least in our experiences thus far, the current state-of-the-art in conventional ARM approaches including those that use the MapReduce paradigm do not scale well to very large datasets. However, approximate approaches with theoretical guarantees have emerged as tenable alternatives [20], which we will explore for our future research.

As indicated in the introduction, our current effort does not surface associations that are necessarily causal. Finding causal associations in biomedicine involves an elaborate set of criteria [8], all of which cannot be checked using data mining approaches. Recent approaches [14], nevertheless, have attempted to identify causal association rules from clinical datasets using automated approaches for identifying confounders [15]. We are currently employing these approaches to identify causal rules for depressive and bipolar disorders. Temporal precedence of the antecedent items to consequents of interest is critical in causality besides accounting for confounders. We will employ advances in ARM that account for temporal sequences [13, 16] in our future work. Finally, we plan to move toward associative classification [32] approaches for specific chronic diseases and for designing new classification features for extracting coded information [10, 11] as a further step from ARM.

Acknowledgments

We are grateful to anonymous reviewers for their helpful comments that improved the presentation of this paper and for interesting suggestions to extend our work using a event sequence framework. This work is supported by the National Center for Advancing Translational Sciences through Grant UL1TR000117 and the Kentucky Lung Cancer Research Program through Grant PO2-415-1400004000-1. The content of this paper is the responsibility of the authors and does not necessarily represent the official views of the NIH.

6. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499. Morgan Kaufmann Publishers Inc., 1994.

[2] S. E. Brossette, A. P. Sprague, J. M. Hardin, K. B. Waites, W. T. Jones, and S. A. Moser. Association rules and data mining in hospital infection control and public health surveillance. *Journal of the American Medical Informatics Association*, 5(4):373–381, 1998.

[3] B. Druss and E. Walker. Mental disorders and medical comorbidity. `http://www.rwjf.org/content/dam/farm/reports/issue_briefs/2011/rwjf69438`.

[4] L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, 38(3):9, 2006.

[5] W. Hämäläinen. Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures. *Knowledge and Information Systems*, 32(2):383–414, 2011.

[6] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *ACM SIGMOD Record*, volume 29, pages 1–12. ACM, 2000.

[7] Healthcare Cost and Utilization Project. Clinical classifications software (CCS) for ICD-9-CM.

https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp.

[8] A. B. Hill. The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58(5):295–300, 1965.

[9] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, Oct. 2002.

[10] R. Kavuluru and Y. Lu. Leveraging output term co-occurrence frequencies and latent associations in predicting medical subject headings. *Data & Knowledge Engineering*, 94(Part B):189–201, 2014.

[11] R. Kavuluru, A. Rios, and Y. Lu. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial intelligence in medicine*, 65(2):155–166, 2015.

[12] R. C. Kessler, P. Berglund, W. T. Chiu, O. Demler, S. Heeringa, E. Hiripi, R. Jin, B.-E. Pennell, E. E. Walters, A. Zaslavsky, and H. Zheng. The US national comorbidity survey replication (NCS-R): design and field procedures. *International Journal of Methods in Psychiatric Research*, 13(2):69–92, 2004.

[13] B. Letham, C. Rudin, and D. Madigan. Sequential event prediction. *Machine learning*, 93(2-3):357–380, 2013.

[14] J. Li, T. D. Le, L. Liu, J. Liu, Z. Jin, B. Sun, and S. Ma. From observational studies to causal rule mining. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):14, 2015.

[15] Y. S. Low, B. Gallego, and N. H. Shah. Comparing high-dimensional confounder control methods for rapid cohort studies from electronic health records. *Journal of comparative effectiveness research*, 5(2):179–192, 2016.

[16] T. H. McCormick, C. Rudin, and D. Madigan. Bayesian hierarchical rule modeling for predicting medical conditions. *Ann. Appl. Stat.*, 6(2):652–668, 06 2012.

[17] S. Moens, E. Aksehirli, and B. Goethals. Frequent itemset mining for big data. In *Big Data, 2013 IEEE International Conf. on*, pages 111–118. IEEE, 2013.

[18] J. A. Morris and M. J. Gardner. Statistics in medicine: Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *British Medical Journal*, 296(6632):1313–1316, 1988.

[19] C. Ordonez, N. Ezquerra, and C. A. Santana. Constraining and summarizing association rules in medical data. *Knowledge and Information Systems*, 9(3):259–283, 2006.

[20] M. Riondato, J. A. DeBrabant, R. Fonseca, and E. Upfal. PARMA: A parallel randomized algorithm for approximate association rules mining in mapreduce. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 85–94. ACM, 2012.

[21] S. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004.

[22] B. Rosner. *Fundamentals of biostatistics*. Cengage Learning, 2015.

[23] E. Ryu, A. M. Chamberlain, R. S. Pendegraft, T. M. Petterson, W. V. Bobo, and J. Pathak. Quantifying the impact of chronic conditions on a diagnosis of major depressive disorder in adults: a cohort study using linked electronic medical records. *BMC psychiatry*, 16(1):1, 2016.

[24] W. R. Shadish, T. D. Cook, and D. T. Campbell. *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Pub., 2002.

[25] N. H. Shah, P. LePendu, A. Bauer-Mehren, Y. T. Ghebremariam, S. V. Iyer, J. Marcus, K. T. Nead, J. P. Cooke, and N. J. Leeper. Proton pump inhibitor usage and the risk of myocardial infarction in the general population. *PLoS One*, 10(6):e0124653, 2015.

[26] I. N. M. Shaharanee, F. Hadzic, and T. S. Dillon. Interestingness measures for association rules based on statistical validity. *Knowledge-Based Systems*, 24(3):386–392, 2011.

[27] P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 32–41. ACM, 2002.

[28] T. Uno, M. Kiyomi, and H. Arimura. LCM Ver.3: Collaboration of array, bitmap and prefix tree for frequent itemset mining. In *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, OSDM '05, pages 77–86. ACM, 2005.

[29] G. I. Webb and J. Vreeken. Efficient discovery of the most interesting associations. *ACM Trans. on Knowledge Discovery from Data*, 8(3):15, 2014.

[30] A. Wright, E. S. Chen, and F. L. Maloney. An automated technique for identifying associations between medications, laboratory results and problems. *J. of Biomedical Informatics*, 43(6):891 – 901, 2010.

[31] A. Wright, A. McCoy, S. Henkin, M. Flaherty, and D. Sittig. Validation of an association rule mining-based method to infer associations between medications and problems. *Applied Clinical Informatics*, 4(1):100, 2013.

[32] X. Yin and J. Han. CPAR: Classification based on predictive association rules. In *SIAM International Conf. on Data Mining*, volume 3, pages 331–335, 2003.

[33] M. J. Zaki. Scalable algorithms for association mining. *Knowledge and Data Engineering, IEEE Transactions on*, 12(3):372–390, 2000.