Reverse Engineering Measures of Clinical Care Quality: Sequential Pattern Mining

Hsuan Chiu[1] and Daniella Meeker[2]

[1] Dept. of Computer Science, University of California, Los Angeles, USA
cherylautumn@cs.ucla.edu
[2] Dept. of Preventive Medicine, University of Southern California, USA
dmeeker@usc.edu

Abstract. Pattern mining has been applied to classification problems. However, detection and analysis of frequently occurring patterns in clinical data is less studied. Instead, data-driven measures of the quality of clinical care are based on abstractions from clinical guidelines, and often are not validated on the basis of outcomes. We hypothesize that by using outcomes as a training signal, we can discover patterns of treatment that lead to better or worse than expected outcomes. Because clinical data is often censored, traditional classification algorithms are inappropriate. In addition, it is difficult to infer the latent meanings of patterns in clinical data if frequency is the only explanation. In this paper, we present a framework for discovering critical patterns in censored data. We evaluate this framework by comparing the patterns we detect with guidelines. Our framework can improve the accuracy in survival analysis and facilitate discovery of patterns of care that improve outcomes.

Keywords: Sequential Pattern Mining, Survival Analysis

1 Introduction

The increased adoption of electronic health records (EHRs) creates new opportunities for both medical discovery and measurement of the quality of care. These activities have largely been conducted in parallel with little dialogue between researchers. The intent of this work is to describe a framework for discovering patterns of care that lead to better or worse than expected outcomes.

1.1 Electronic Clinical Quality Measures

The widespread adoption of electronic health records has led to a corresponding interest in using information capture in these applications to measure the quality of patient care. The Federal Meaningful Use program is an incentive program to healthcare providers and hospitals to encourage adoption of EHRs. They have released over 100 such measures that have been developed by expert consensus. These measures operationalize clinical guidelines into metrics that are based on transactional data elements

in structured fields in the EHRs. These electronic clinical quality measures (eCQMs) are generally reported as the ratio of patients that have received high quality of care to those that have not given that a guideline applies. If a guideline is applicable to a particular patient he is in the "denominator" of the ratio- the inclusion and exclusion criteria for measure denominators are defined on the basis of EHR documentation. The measure numerator is usually either a clinical outcome (e.g. lab test results indicate good control of diabetes) or a clinical process (treatment with beta blockers). Like most metrics from transactional databases, the data elements in a quality measure come in the form of time-stamped events – most commonly the time stamp is a data entry event, but potentially also a reported time of an event, as in the history of a heart attack.

While some eCQMs are based on very complicated patterns that define a sequence of inclusion and exclusion criteria intended to increase the specificity of the applicable populations, some guidelines that have been operationalized into electronic quality measures are more straightforward. Furthermore, many quality measures that are abstracted from clinical guidelines have not been shown to have predictive validity on the basis of certain outcomes [1]. These characteristics make many eCQMs poor candidates for evaluating our framework, but also demonstrate the need for a framework that links process of care to outcomes. Some of the simplest eCQMs are represented by known Drug-Drug-Interactions (DDIs) – patients concurrently exposed to two drugs known to interact are likely to have worse outcomes than similar patients that were not exposed to both drugs. The knowledgebase of DDI is standardized and well-developed, with the denominator population easily defined on the basis of exposure to one of two drugs known to interact. For these reasons, we can assess the face validity of our results on the basis of our ability to detect DDIs.

1.2 Methods for Mining Clinical Data

Large observational datasets provide a valuable compliment to the gold standard of randomized clinical trials. It has often been pointed out that clinical trials have a careful selection of uncomplicated subjects that may not reflect real-world exposures [2][3]. A less frequently acknowledged value conferred by observational analyses is a richer complexity of treatment histories and combinations of exposures to multiple therapies. Therefore, some data mining approaches like text mining [4], temporal pattern mining [5], [6] or sequential pattern mining (SPM) [7], [8] have been applied to medical data with the expectation of accelerating novel knowledge discovery. However, some new issues are raised after these general data mining techniques are directly applied to medical data.

Data mining in medicine is differentiated from other fields insofar as the notion of "comprehensibility" plays an important role [9], and hypothesis-generating studies such as these must also have external validity and comport with clinical models. Because observational studies cannot control for selection bias, they must be conducted and interpreted carefully for purposes of causal inference [10]. Therefore, one big issue raised from data mining results, such as sequential pattern mining, is the interpretation difficulty because results are mined according to frequency, which not only

generates too many similar patterns but also hard to explain these patterns' latent meaning.

Outside of biomedical literature, sequential pattern mining has been more broadly adopted to solve classification problems. For example, Cheng et al. [11], [12] applied the discriminative pattern mining on software failure detection and trajectories on road network classification. While classifiers, such as logistic regression or SVM [13] are commonly used tools in biomedical literature and data mining, important methods in a clinical data analyst's arsenal are survival analyses.

Survival analysis accounts for censoring – the lack of complete follow up on outcomes used to train classification algorithms. Most clinical data where outcomes can only be observed after extended time has elapsed have this limitation. Censoring means the precise survival time cannot be fully captured by the observational data. For example, suppose two patients who entered our dataset at age 90, one of which was observed for 8 years before he died and the other was observed for 6 months before becoming lost to follow up. At the end of the study, we do not know the actual survival time of the patient followed up to 6 months because his outcome, death, is not observed. In terms of modeling censored data, survival analysis is better than classification because survival analysis accounts not only for the likelihood of outcomes and exposure of interest, but also for the likelihood that each subject could have been observed given the observation length. Using binary classification to model censored data has several drawbacks. First, we cannot simply classify patients as "alive" or "dead" because some actual outcomes may not be observed. Second, the number of observed events is typically significantly undersized in the population, leading to a skewed dataset.

So far, discovering critical patterns in censored data is less studied. Mining critical patterns help researchers identify which patterns play key roles in the survival probability. For purposes of causal inference, it is also important to interpret the latent meanings of patterns, such as their relative influence upon the survival probability.

Considering the problems mentioned above, we will solve two problems in this paper. First, we discover a set of critical sequential patterns, which have stronger relationships with the survival outcome after an incident diagnosis from the censored data. For example, if we incorporate these patterns as covariates into survival analysis, such as *Cox proportional hazard regression model* [14], these patterns should perform as reliable predictors. Secondly, we expect the latent meanings of these critical patterns to be interpretable, such as to what extent these critical patterns influence the survival outcome. To the best of our knowledge, mining the critical sequential patterns in the censoring data for survival analysis is are yet to be studied.

The rest of the paper is structured as follows: In Section 2, we discuss related work about data mining in health data. Our framework about how to discover reliable frequent patterns as covariates is introduced in Section 3. In Section 4, we present the experimental evaluations. Lastly, we conclude the paper with study limitation and future work in Section 5.

2 Related Work

2.1 Care pathways, treatment patterns, and outcomes in healthcare databases

Treatment guidelines and care pathways are currently developed primarily by deliberative expert consensus to promote the practice of "evidence based medicine". The evidence under consideration is typically in the form of a systematic review of the literature, with higher 'evidence value' being placed on randomized clinical trials that may or may not have ecological validity. These care pathways are often operationalized into quality indicators and performance metrics as process measures that inform policy and, in turn, practice. Despite this careful attention to evidence in the literature, and acknowledgement of the importance of predictive validity by organizations such as the National Quality Forum, there have been few studies investigating whether pathways and patterns in quality indicators are indeed associated with better outcomes in the real world after adjusting for underlying risk factors. We argue that there is not only a need to bolster the evidence that "evidence-based medicine" is effective, but also that there may be undiscovered patterns and pathways that lead to *better than expected* outcomes that exploratory analysis might surface as candidates for quality indicators.

By contrast, there has been extensive attention to *worse than expected* outcomes for drug treatments in the field of pharmacovigilance[15]. These studies, while originally focused on adverse event reporting databases, soon extended to include analysis of the same data sources that are being used to compute performance metrics – administrative claims and electronic medical records. These studies have fallen into two categories – (1) post-market surveillance in the form of risk-adjusted hypothesis testing (with a focus on specific drugs and outcomes) and (2) exploratory data mining to potentially identify new patterns. While the first category of work has generated substantial innovations in risk-adjustment methods that are relevant to addressing selection bias [16] (a primary limitation of observational data analysis), we are more interested in the second category for the purposes of this work. In addition to conventional association mining between a single drug and a single outcome in the form of disproportionality analysis, there have been studies that have explored combinations and sequences to generate new hypotheses about combinations of drugs [17] and methods for detecting interesting temporal patterns [18]. Similar to genome-wide association studies, this type of exploratory analysis requires insuring against spurious correlations and multiple comparisons. [9] [19]

2.2 Pattern mining in medical domain

Currently, applying data mining to medical data is a growing trend and most data mining applications in medical fields are directly using the state of art approaches like classical classifier, clustering or association rules to derive results [20]. However, directly applying these generalized methods to health data still cannot achieve satisfied expectations. In health studies, survival analysis is one of the most important statistical approaches. Since the health data is often censored due to the termination of

a study or the failure to follow-up observation subject, usually the outcome of interest in survival analysis is the time-to-event data. Thus far, there are only a few works studied in the relationship between frequent sequential treatment patterns and survival time. Silva et al [8] studied how to evaluate the relationship between survival time with sequential treatment patterns. They used Kaplan Meier [21] to estimate the median survival time among a set of patients who have the same treatment patterns and further pruned out patterns with shorter median survival time. In our work, we further examine how each sequential treatment pattern will influence the survival probability. Malhotra et al [7] also used a sequential pattern mining technique to retrieve frequent sequential treatment patterns for Glioblastoma Multiforme (GBM). They formulate their problem as a classification problem, and use these sequential treatment patterns as additional features to predict whether a patient can survive longer than the median survival period or not. In our work, we consider using Cox proportional hazard regression instead of classification to model the survival problem since most health data contains censored issues.

3 Framework

3.1 Data Description

Source of Data. The primary source of data for this study was administrative claims submitted to insurance companies by healthcare providers to receive reimbursement for services. Administrative claims lack the clinical detail that might be present in electronic health records, but have the benefit of capturing care and outcomes across all of the healthcare providers from whom a patient has received care. These administrative data sets were aggregated and cleaned by the Innovation in Medical Evidence Development and Surveillances (IMEDS) lab [22] hosted by Reagan-Udall Foundation for the FDA. This clinical data is translated into standardized vocabularies containing all of the medical code sets, terminologies, vocabularies and ontologies taxonomies. Drugs are also coded with RxNorm, which is a drug reference terminology maintained by the National Library of Medicine (NLM), and conforms to the Observational Medical Outcomes Partnership (OMOP) Common Data Model originally developed by the Foundation for the National Institutes of Health (http://omop.fnih.org).

Clinical Population. We randomly sampled 42,365 patients from a total of 1,027,339 patients diagnosed with Congestive Heart Failure (CHF) between January 1, 2003 and March 31, 2003. Among these 42,365 patients, 1,599 death events were observed. We used gender, Deyo's Charlson Comorbidity Index Variables, [23] and the age at CHF index diagnosis date as a part of patient features. Summary statistics are listed in Table 1. We used random samples of 42,365 patients to run the experiments due to computational constraints. This approach is valid for the following reasons: 1) We wished to verify whether frequent patterns can perform as reliable predictors in censored data, and 2) We want to ascertain whether the latent meaning of frequent patterns can be

discovered through our method. Our experiments show that our framework has potential for achieving these goals.

Table 1. Population description.

Covariate	Proportion (%)
Male	54.04
Myocardial Infarction	8.90
Peripheral Vascular Disease	3.05
Cerebrovascular Disease	12.27
Dementia	0.28
Chronic Pulmonary Disease	27.53
Rheumatologic Disease	4.24
Peptic Ulcer Disease	2.11
Mild Liver Disease	1.75
Diabetes	33.42
Diabetes with Chronic Complications	9.10
Hemiplegia or Paraplegia	1.07
Renal Disease	10.08
Moderate or Severe Liver Disease	5.86
AIDS	0.43

Covariate	Min	Max	Mean	Std
Age on Index Date (years)	0	89	53	11.5

3.2 Framework Overview

We briefly introduce our framework and then illustrate details in each section. Initially, we set a censored date and randomly sample a set of patients with demographic, disease and drug information. Then, we construct two types of features for each patient. The first are baseline characteristics composed of demographic and health status at index diagnosis, and another is the treatment feature type derived from sequential pattern mining. Because the quantity of treatment features may grow to more than 5,000, when we relax the support threshold, we screen out those inactive features before we incorporate them into Cox regression model. Thus, we take two screening strategies to do the ranking. We select only top-K treatments after controlling for baseline health status and age in the Cox model. We hypothesize that the screening strategy can reliably bring predictive patterns into the model.

3.3 Feature generation

In our study, we generate two types of feature for patients. The first are baseline characteristics that cannot be changed during the course of medical care, such as age at

index diagnosis, sex and comorbidity diseases listed in **Table 1**. For each patient, the first date that s/he was diagnosed with CHF is referred to as index date. The comorbid disease features are based on Charlson Comorbidity Index Indicators [23] - 13 diseases coded by ICD9 (International Classification of Diseases and Related Health Problem, v9) such as renal disease, liver disease and HIV. For example, if a CHF patient also has liver disease and heart disease before the index date, we will assign a binary value 1 in these two features. By controlling for these covariates we will detect patterns that arise independent of health status.

The covariates of interest in this work (treatment covariates) are the treatment sequence after the index date. We build treatment covariates by using sequential pattern mining to extract patterns from patients' drug history after the index date. We view patients' drug history as a transaction database D which contains a set of tuples (pid, tid, Itemset), where pid is a patient id, tid is a transaction id based on the prescription time, and Itemset is a set of drugs prescribed on the same day. All these tuples with the same pid can be regarded as a sequence of itemsets ordered by increasing tid. Thus, we can leverage the state-of-the-art sequential pattern mining algorithm to generate a set of frequent sequential patterns as treatment patterns.

In this work, we leverage SPADE[24] and VMSP[25] as our sequential pattern mining approach. The difference between SPADE and VMSP is that the former generates a complete set of frequent patterns while the latter generates maximum length frequent patterns. Treatment covariates are denoted as binary value. If a frequent pattern occurs in a patient's sequence, we assign it as true; otherwise we label it as false.

3.4 Feature Screening

So far, we generate a feature set including patients' demographic information, given health conditions and a set of frequent sequential treatment patterns. When the number of patterns is massive, only critical patterns are necessary for building a regression model. Most regression models leverage feature selection strategy by adding a specific penalty function into the regression model. When the quantity of covariates and the sample size are large, high computational cost makes this method inefficient. In this work, we are focusing on how to effectively select critical patterns before we use them to train a regression model. We also do not want to limit the feature selection strategy for a certain type regression model. Here we provide two approaches to achieve pruning.

Model-free screening. To estimate the discriminative power of a feature, we leverage the novel feature screening method proposed by Zhu et al [26]. The most distinguishable point of their method, Model-free screening (MFS), is that the ranking procedure widely covers many parametric and semi-parametric models, such as linear regression, logistic regression and Cox proportional hazard regression. When the number of covariates is huge and the information about the underlying model is limited, MFS provides a great flexibility to rank these features.

In MFS, covariates $\mathbb{Z} = (Z_1, Z_2, \dots, Z_p)^T$ are classified into two sets, active predictors with non-zero coefficient, $\mathbb{Z}_A$, and inactive predictors with zero coefficients $\mathbb{Z}_I$. Their method claims that $\mathbb{Z}_A$ can be consistently ranked before $\mathbb{Z}_I$.

Given a sample with size n, they use $\widehat{w_k}$ as a natural estimator for measuring the marginal utility of the k'th element in $\mathbb{Z}$, where

$$\widehat{w_k} = \frac{1}{n}\sum_{j=1}^{n}\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{Z}_{ik}I(Y_i < Y_j)\}^2$$

with assumptions $\frac{1}{n}\sum_{i=1}^{n}\mathbb{Z}_{ik} = 0$ and $\frac{1}{n}\sum_{i=1}^{n}\mathbb{Z}_{ik}^2 = 1$. By ranking $\widehat{w_k}$ in descending order, their approach is claimed to consistently screen out inactive predictors. To implement their ranking method, the first step is to rank tuples in the sample set according to the outcome variable Y in ascending order. Next, calculating $\widehat{w_k}$ for each covariate is through scanning $\mathbb{Z}_k$ value in n tuples. Lastly, order the p covariates by $\widehat{w_k}$ in descending sequence. The complexity requires $O(nlogn + np + plogp)$.

Maximized coverage screening. We can observe that when the sample size n goes large, the complexity of MFS will be dominated by $O(np)$. In this work, we provide another heuristic strategy to rank these features in a more efficient way.

Our basic idea is to choose the feature with maximized coverage of data points in each round. At the first step, we rank sequential patterns by supports in descending order. Then, we pick the one with the widest coverage rate, which means the largest number of individuals who own this feature. If more than one feature contain the same highest coverage rate, we choose the feature with the highest support. After we select a feature, all individuals having this feature will be eliminated. In the next iteration, the identical choosing strategy is applied, but to a smaller set of individuals. When all individuals are eliminated, we recover all individuals back and we continue the selection procedure until all features are ranked.

In order to efficiently select the feature with the highest coverage rate, we use the bitmap representation to denote how a sequential pattern distributes among individuals in the sample dataset. Initially, we have an empty bitmap, BM_{all}, with length $|BM_{all}| = n$ and all bits in the BM_{all} are set to zero. For each sequential pattern i, a bitmap BM_i is created and the index of each bit in a bitmap represents each individual in the dataset. If a sequential pattern *i* occurs in individual *j*'s sequence, the *j*'th bit will be set to one in *i*'s bitmap, $BM_i(j) = 1$; otherwise, $BM_i(j) = 0$. We apply BM_i ANDNOT BM_{all}, and then counting the cardinality of BM_i, meaning the number of new individuals are covered by the feature i in current iteration. After the highest coverage sequential pattern h derived in current iteration, we simply use BM_hOR BM_{all} to represent how many individuals are already covered so far. Then in the next

round, we use the updated BM_{all} to discover the next candidate among the rest sequential patterns. Once all bits in BM_{all} are set to one, we reset the BM_{all}. The ranking process is continued until all sequential patterns are ranked. Since computing the coverage rate can be regarded as constant time by implementing in bitmap, the total cost is bounded by $O(p^2)$.

ALGORITHM: Maximize Coverage Screening

```
Input: Frequent sequential_pattern_set F
Output: An ordered selected pattern set F_s
Sort patterns in F according to support in descending order.
currentCoverage←0
maxCoverage←0
BM_all ←clear each bits
BM_max ←clear each bits
While F is not null
  For each pattern p in F, do
     BM_p.ANDNOT(BM_all)
      currentCoverage← |BM_p|
      if currentCoverage > maxCoverage
        maxCoverage=currentCoverage;
        BM_max = BM_p
  if BM_max is not null
     BM_all.OR(BM_max)
     F.remove(pattern max)
     F_s.add(pattern max)
  else BM_all ←clear each bits
end
return F_s
```

3.5 Feature Construction

In our study, we leverage both immutable covariates and treatment covariates as features in the Cox model. After we rank treatment covariates either by MFS or by Maximized Coverage, only top-K patterns will be incorporated into the regression model. The reason that we need to include immutable covariates is we do not want to incorrectly assume people are dying because of the drug rather than the illness. We set a censored date as the termination of the observation, and we set the event as death. Finally, we leverage immutable covariates, treatment covariates and outcomes to train the Cox model.

4 Experimental Evaluation

In this section, we evaluate the performance of the Cox model incorporated with frequent patterns. All experiments are conducted on 8 EC2 Compute Units (4 virtual cores with 2 EC2 Compute Units each) with 15GB of main memory, running on Linux 64-bit platform.

Our goal in this study is to select critical sequential patterns as covariates in the Cox regression adjusted for underlying risk. We expect that the selected sequential patterns can be sufficient enough to perform as reliable predictors in the Cox regression. In our evaluation, as the response variable of Cox regression contains both binary and continuous variable which associate with time, we chose to use RisksetAUC[27] as our main measurement.

4.1 Approaches for comparison

We compare three different ranking mechanisms to order sequential patterns and we apply the greedy forward selection to iteratively pick top-K sequential patterns as Cox regression covariates. For each Cox regression, we measure the RisksetAUC value as the performance. The benchmark ranking method is to order sequential patterns by the support value from high to low. The other two ranking methods are MFS and Maximized coverage.

The experiment data is extracted randomly from IMEDS with medical records ranging from January 1, 2003 to March 31, 2013 among 1,027,339 CHF patients. Initially we set censored date on October 10, 2014 and we randomly selected 42,365 patients. Next, we apply both SPADE and VMSP in package spmf [28], with setting minimum support threshold 0.1, 0.05 and 0.025. Table 2 shows the number of sequential patterns generated in dataset DS2014.

Table 2. The number of features generated.

Method	SPADE			VMSP		
Threshold	0.1	0.05	0.025	0.1	0.05	0.025
Number of Patterns	85	643	5162	59	373	2,750

4.2 Results

Accuracy of Survival Analysis controlling for Charlson Index indicators in absence of treatment covariates. Table 3 shows coefficients of Cox model trained with patients' immutable features in dataset DS2014. Positive coefficient represents a patient that has that feature, and his hazard ratio is expected to increase. We can see that most coefficients are positive and this result is reasonable since comorbid diseases denoted in Charlson index indeed have the potential to increase the risk among each other.

Table 3. Coefficients and accuracy of immutable features.

	coef	p-value
Gender	0.002	0.278
Myocardial Infarction	0.205	0.013
Peripheral Vascular Disease	0.050	0.699
Cerebrovascular Disease	-0.004	0.958
Dementia	-0.231	0.609
Chronic Pulmonary Disease	0.369	0
Rheumatologic Disease	0.313	0.002
Peptic Ulcer Disease	0.308	0.019
Mild Liver Disease	1.070	0
Diabetes	0.058	0.323
Diabetes with Chronic Complications	0.217	0.008
Hemiplegia or Paraplegia	0.926	0
Renal Disease	0.884	0
Moderate or Severe Liver Disease	0.619	0
AIDS	0.949	0
Age on Index Date	0.035	0
AUC		0.648

Accuracy Results. Since the feature set of the Cox model can be formed by three ranking mechanisms, (support, MFS and Maximized coverage), we compare how these ranking methods affect the Cox model accuracy when we incorporate top-K features into the model. In each figure, we list the total number of patterns generated by the pattern mining algorithm associated with a threshold. For example, the main title, SPADE_0.1, in Figure 1a means we apply SPADE with threshold 0.1 to generate frequent patterns and we obtain total 85 patterns in this case. The accuracy depicted in all figures starts from the base AUC, 0.648.

In some cases, including all patterns into the model is allowable, such as **Figure 1**a and **Figure 1**d. However, in some other cases, not all patterns are suitable to be fully incorporated due to the large number of patterns, such as **Figure 1**c and **Figure 1**f. Thus, the feature screen strategy is important in the latter scenarios. For example, in **Figure 1**c, if we select only 10% of features into the Cox model and attain higher than 85% accuracy, this implies that these features are sufficient to play as reliable predictors for the dataset.

From **Figure 1** we can observe that the base accuracy is enhanced significantly when we bring these patterns into the model initially, and then the accuracy grows stable until all features are included. Compare to MFS and Maximized coverage, we also observe that the support ranking strategy does not effectively choose reliable predictors into top-K feature set in most cases because the RisksetAUC of support does not show competitive accuracy until we include all features into the Cox model, such as **Figure 1**a and **Figure 1**d. This implies that patterns chosen by support may not be sufficient enough to explain the outcome unless we apply all patterns to describe it. Next, we can view that features filtered by MFS provide highest accuracy when we include only a few of them into the model. When we incorporate more covariates, Maximized coverage is able to support equal or higher discriminative ability.

This observation delivers flexibility of choosing the ranking strategy from MFS or Maximized coverage. For example, if we want to preserve just a few features, such as the number is less than 100, MFS will be a good option for feature screening. However, if we want to keep more features at the beginning because the size of frequent pattern is large (such as more than 5000), Maximized Coverage serves as a better choice. The reason is that Maximized Coverage not only saves more computation time, especially when the number of features and the number of tuples are massive, but also attains equal or higher accuracy than MFS.

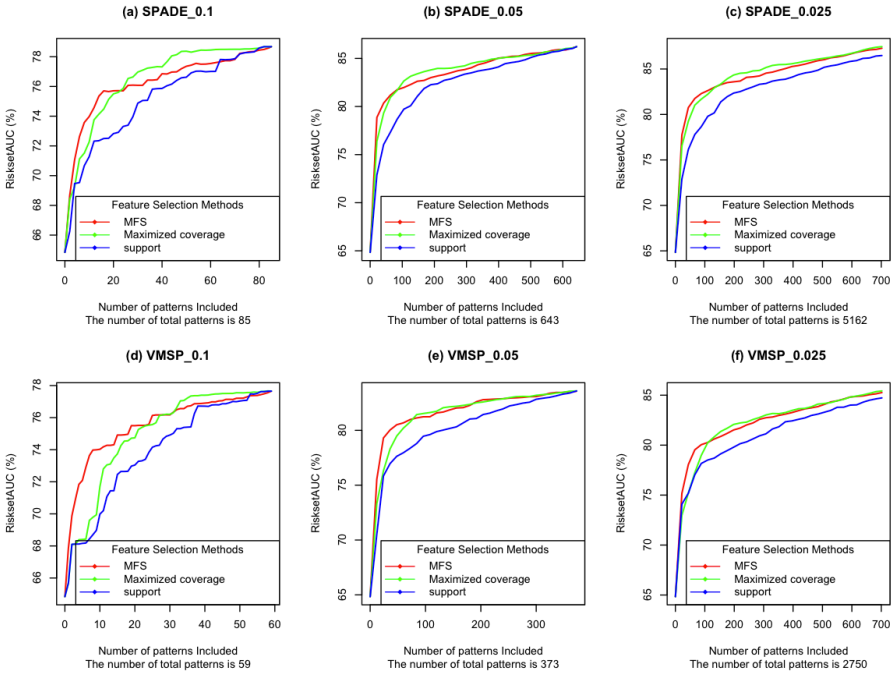

Figure 1. The effect of ranking mechanism on accuracy for DS2014.

Empirical Validation. In this section, we want to ascertain the validity of our selected critical patterns, and we use drug-drug interaction (DDI) as our sanity check. We run a simple validation as following. For each Cox model, we apply Benjamin Hochberg method to adjust each feature's p-value for reducing false discovery rate, and then we select positive coefficient feature with adjusted p-value less than 0.05. The latent meaning of the selected feature is: 1) this feature is significantly related to the outcome, and 2) if a patient has this feature, his hazard ratio is expected to increase. Next, we use external knowledgebase[29] to see whether this selected pattern contains known DDI. The external knowledgebase classifies DDI into two levels, moderate and severe. If the selected pattern contains DDI, we list it in the results.

We use Cox models trained from DS2014 to run the DDI validation. Frequent patterns are generated by VMSP with threshold 0.05, and patterns are ordered by sup-

port, MFS and Maximized Coverage, respectively. Due to space constraints, we only list distinct patterns detected with DDI in Table 4. For example, in Table 4, when the Cox model incorporating top 120 patterns, ranked by MFS, as covariates, the pattern {Acetaminophen → Warfarin} is significant and this pattern with positive coefficient is also verified as having severe DDI. We observe that our proposed framework has potential to select truly critical patterns and some of their latent meanings can be verified. For the other significant patterns, they might be candidates in new medical knowledge discovery after further investigation with medical domain experts.

Table 4. Verified DDI patterns mined by VMSP

	PtnN	DDI	coef	adjP	Drug sequence
support	48	Moderate	0.353	0.001	Furosemide → Albuterol
	84	Moderate	0.450	0.000	salmeterol fluticasone
	132	Moderate	0.343	0.049	Azithromycin → Prednisone
	228	Severe	0.350	0.044	Acetaminophen → Warfarin
	348	Moderate	0.408	0.041	Lisinopril → Furosemide
MFS	120	Severe	0.349	0.001	Acetaminophen → Warfarin
	132	Severe	0.324	0.018	Furosemide → Warfarin
	288	Moderate	0.384	0.040	Albuterol → Prednisone
	300	Severe	0.384	0.045	Azithromycin → Levofloxacin
	312	Moderate	0.520	0.000	salmeterol fluticasone
	336	Moderate	0.396	0.047	Lisinopril → Furosemide
Maximized Coverage	60	Moderate	0.216	0.039	salmeterol fluticasone
	72	Moderate	0.303	0.008	Lisinopril → Furosemide
	96	Moderate	0.228	0.049	Furosemide → Albuterol
	132	Moderate	0.351	0.023	Albuterol → Prednisone

5 Conclusion

In this paper, we proposed a framework to efficiently discover critical sequential patterns as reliable predictors in censored data. We used Cox regression to model the censored data in order to avoid the skewed data problem in classification. We applied SPM to generate frequent patterns and we provide two feature-ranking methods, MFS and Maximized coverage, to select important patterns. Our main contribution is to accurately screen out those insignificant but frequent sequential patterns. In experiments, we demonstrated that the discovered sequential patterns are able to improve the prediction accuracy, and we also showed that the Cox regression model provides a better way to explain the latent meaning of sequential patterns.

The framework we adopted has some limitations that might be addressed in future work. By adopting an epidemiological framework that holds fixed the disease state at the index date of CHF diagnosis, we are not addressing the dynamic nature of disease evolution and the challenges associated multi-morbid patients and complex interactions between disease, side-effects, and treatments. In future work, we might consider Bayesian networks to help disentangle some of these confounding effects and incor-

porate knowledge that has been generated in hypothesis-driven research. This approach would help resolve confounding concerns between drugs used to treat conditions that evolve after the index date and indications that may be side-effects.

References

1. J. Mant, "Process versus outcome indicators in the assessment of quality of health care," *Int. J. Qual. Heal. Care*, 2001.
2. D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," *Artif. Intell. Med.*, vol. 34, no. 2, pp. 113–127, 2005.
3. R. S. D'Agostinio and R. J. D'Agostinio, "Estimating treatment effects using observational data.," *J. Am. Med. Assoc.*, vol. 297, no. 3, pp. 314–316, 2007.
4. K. D. Shetty and S. R. Dalal, "Using information mining of the medical literature to improve drug safety.," *J. Am. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 668–674, 2011.
5. G. N. Nor, A. Bate, J. Hopstadius, K. Star, and I. R. Edwards, "Temporal Pattern Discovery for Trends and Transient Effects : Its Application to Patient Records," pp. 963–971, 2008.
6. G. N. Norén, J. Hopstadius, A. Bate, K. Star, and I. R. Edwards, "Temporal pattern discovery in longitudinal electronic patient records," pp. 361–387, 2010.
7. K. Malhotra, D. H. Chau, J. Sun, C. Hadjipanayis, and S. B. Navathe, "Temporal Event Sequence Mining for Glioblastoma Survival Prediction," in *KDD 2014 Workshop on Health Informatics (HI-KDD 2014)*, 2014.
8. A. Silva, W. M. Jr, O. Queiroz, and M. Cherchiglia, "Sequential Medical Treatment Mining for Survival Analysis," in *SBBD*, 2009, pp. 166–180.
9. R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," *International Journal of Medical Informatics*, vol. 77, no. 2. pp. 81–97, 2008.
10. E. S. Fisher, D. E. Wennberg, D. A. Alter, and M. J. Vermeulen, "Analysis of observational studies in the presences of treatment selection bias: Effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods," *J. Am. Med. Assoc.*, vol. 297, no. 3, pp. 278–285, 2007.
11. J. G. Lee, J. Han, X. Li, and H. Cheng, "Mining discriminative patterns for classifying trajectories on road networks," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 5, pp. 712–726, 2011.
12. C. Lo, D. and Cheng, H. and Han, J. and Khoo, S.C. and Sun, "Classification of Software Behaviors for Failure Detection : A Discriminative Pattern Mining Approach," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining}*, 2009, vol. 73, no. 2, pp. 557–565.
13. C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.
14. D. R. Cox, "Regression models and life tables," *J. R. Stat. Soc. Ser. B*, vol. 34, no. 2, pp. 187–220, 1972.

15. M. Liu, M. E. Matheny, Y. Hu, and H. Xu, “Data Mining Methodologies for Pharmacovigilance,” *SIGKDD Explor. Newsl.*, vol. 14, pp. 35–42, 2012.
16. S. Schneeweiss, J. A. Rassen, R. J. Glynn, J. Avorn, H. Mogun, and M. A. Brookhart, “High-dimensional propensity score adjustment in studies of treatment effects using health care claims data.,” *Epidemiology*, vol. 20, no. 4, pp. 512–522, 2009.
17. N. P. Tatonetti, J. C. Denny, S. N. Murphy, G. H. Fernald, G. Krishnan, V. Castro, P. Yue, P. S. Tsau, I. Kohane, D. M. Roden, and R. B. Altman, “Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels.,” *Clin. Pharmacol. Ther.*, vol. 90, no. 1, pp. 133–142, 2011.
18. H. Jin, J. Chen, H. He, G. J. Williams, C. Kelman, and C. M. O’Keefe, “Mining unexpected temporal associations: Applications in detecting adverse drug reactions,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 12, no. 4, pp. 488–500, 2008.
19. L. J. Mazlack, “Using association rules without understanding their underlying causality reduces their decision value,” 2004, pp. 312–319.
20. N. Esfandiari, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, “Knowledge discovery in medicine: Current issue and future trend,” *Expert Syst. Appl.*, vol. 41, no. 9, pp. 4434–4463, Jul. 2014.
21. D. G. D. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text, Third Edition (Statistics for Biology and Health)*. 2011.
22. IMED. http://imeds.reaganudall.org/.
23. R. A. Deyo, D. C. Cherkin, and M. A. Ciol, “Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases.,” *J. Clin. Epidemiol.*, vol. 45, no. 6, pp. 613–619, 1992.
24. M. Zaki, “SPADE: An efficient algorithm for mining frequent sequences,” *Mach. Learn.*, vol. 42, no. 1, pp. 31–60, 2001.
25. P. Fournier-Viger, C.-W. Wu, A. Gomariz, and V. Tseng, “VMSP: Efficient Vertical Mining of Maximal Sequential Patterns,” in *Advances in Artificial Intelligence*, M. Sokolova and P. van Beek, Eds. Springer International Publishing, 2014, pp. 83–94.
26. L.-P. Zhu, L. Li, R. Li, and L.-X. Zhu, “Model-Free Feature Screening for Ultrahigh-Dimensional Data,” *J. Am. Stat. Assoc.*, vol. 106, no. 496, pp. 1464–1475, 2011.
27. P. J. Heagerty and Y. Zheng, “Survival model predictive accuracy and ROC curves 1,” *Biometrics*, vol. 61, no. March, pp. 92–105, 2005.
28. “spmf. An Open-Source Data Mining Library. http://www.philippe-fournier-viger.com/spmf.” .
29. “Healthline. http://www.healthline.com/druginteractions?” .