Integrating Temporal Pattern Mining in Ischemic Stroke Prediction and Treatment Pathway Discovery for Atrial Fibrillation

Shijing Guo, PhD[1], Xiang Li, PhD[1], Haifeng Liu, PhD[1], Ping Zhang, PhD[2], Xin Du, MD[3], Guotong Xie, PhD[1], Fei Wang, PhD[4]

[1]IBM Research - China, Beijing, China
[2]IBM T.J. Watson Research Center, New York, USA
[3]Department of Cardiology, Beijing Anzhen Hospital, Beijing, China
[4]Department of Healthcare Policy and Research, Cornell University, New York, USA

Abstract

Atrial fibrillation (AF) is associated with an increased risk of acute ischemic stroke (AIS). Accurately predicting AIS and planning effective treatment pathways for AIS prevention are crucial for AF patients. Because of the temporality of patients' disease progressions, sequential disease and treatment patterns have the potential to improve risk prediction performance and contribute to effective treatment pathways. This paper integrates temporal pattern mining into the AF study of AIS prediction and treatment pathway discovery. We combine temporal pattern mining with feature selection to identify temporal risk factors that have predictive ability, and integrate temporal pattern mining with treatment efficacy analysis to discover temporal treatment patterns that are statistically effective. Results show that our approach has identified new potential temporal risk factors for AIS that can improve the prediction performance, and has discovered treatment pathway patterns that are statistically effective to prevent AIS for AF patients.

Introduction

Atrial fibrillation, also called AFib or AF, is a common cardiac rhythm disturbance. Over 2.7 million Americans are living with AF[1], and the number even reaches up to approximately 10 million in China[2]. AF patients have a 5-fold increased risk of acute ischemic stroke (AIS)[1]. The condition contributes a double risk of hospitalization with a 2.1% death rate during the hospitalization[1, 3]. Therefore, accurately predicting the risk of AIS and planning effective treatment pathways to prevent AIS occurrence could largely benefit to AF patients.

However, AIS prediction and prevention for AF patients face many difficulties. Current widely-used AIS risk prediction models including Framingham[4] and CHA_2DS_2-VASc[5] have only moderate prediction performance, especially for the short-term (e.g., 1 year) AIS prediction tasks. One main reason of the underperformed results is that the risk factors which the models consider, including age, gender, prior stroke, hypertension, diabetes, etc., are quite limited. To address this problem, our previous works have identified some new potential risk factors[6] of 2-year ischemic stroke from Chinese Atrial Fibrillation Registry (CAFR) data, and also have constructed some interacted risk factors[7] to improve prediction performance. However, all these factors are defined at single time points (e.g., at the baseline of CAFR study). The states of a disease change over time, and factors based on single-time points may be insufficient or incomplete in predicting the AIS risk. The temporal disease progressions in patient medical history could also influence the future AIS occurrence, but have not considered in the previous AIS prediction models.

Moreover, although some interventions, such as warfarin, have been proved to provide effective prevention of AIS in general[8], their efficacy under complex clinical scenarios, e.g., the efficacy of continuous/discontinuous medication usage, or the efficacy for patients with specific disease and treatment histories, are still not completely clear. Without considering the temporal disease and treatment progressions, it is difficult to define a reasonable personalized treatment plans for specific patients.

In order to solve the above difficulties, effective temporal disease and treatment patterns can be identified and be further applied to the relevant studies to improve the performance of the AIS prediction and to discover potential effective treatment pathways to prevent AIS. In data mining, temporal pattern mining methods are widely used to identify the frequently appeared temporal patterns of ordered events. They have been applied in many clinical studies, including identifying meaningful temporal patterns from health records[9] or discovering signature patterns from medical events[10]. Several state-of-the-art methods can be implemented to discover sequential patterns, including GSP[11], sequential pattern discovery using equivalence classes (SPADE)[12], Pre-fixSpan[13], and sequential pattern mining (SPAM)[14]. Some methods can also help to discover abstracted temporal patterns or patterns with time-intervals, which include knowledge-based temporal-abstraction (KBTA)[15], KarmaLego[9], temporal skeletonization[16], and temporal graph mining[10].

Nevertheless, it is still challenging to apply temporal pattern mining in AIS prediction and treatment pathway discovery for AF patients. Because of the pattern explosion problem[16], a large number of temporal patterns with high redundancy are usually identified by the temporal pattern mining methods, which brings the following problems:

1. For risk prediction, not all frequent temporal patterns are risk factors that have predictive ability for the specific outcomes, and the high redundancy may even negatively affect the prediction performance.
2. For treatment pathway discovery, the efficacy of each frequent treatment patterns cannot be ensured by the mining methods, so many of the discovered treatment patterns may not have clinical meanings for AIS prevention.

In order to address these issues, this paper demonstrates the approach of applying frequent temporal pattern mining methods in AIS prediction as well as in treatment pathway discovery for AF patients based on CAFR data. During the process of AIS prediction, we integrate sequential pattern mining (SPAM) with feature selection methods to automatically identify the temporal patterns of disease progressions that have predictive ability to predict AIS as temporal risk factors, and build AIS prediction models based on the combination of selected baseline risk factors and temporal risk factors. In the study of exploring effective treatment pathways, we combine SPAM with treatment efficacy analysis to discover temporal treatment pathways that are statistically effective to prevent AIS from a large number of frequent patterns. The control group and confounding factors are taken into account when we evaluate the efficacy of a treatment pattern. The results show that our approach has identified new potential temporal risk factors of AIS and has improved the performance of AIS prediction for AF patients. Also, it helps to discover potential treatment pathway patterns that are statistically effective for preventing AIS, which could contribute in making personalized care plans for AF patients.

Methods

Two case studies are conducted and demonstrated in the paper. The first study aims to build one-year AIS prediction models for AF patients, based on patient medical history data. The second study aims to discover potential significantly effective treatment pathways of one-year patient treatments.

All studies in this paper are based on Chinese Atrial Fibrillation Registry (CAFR)[2] data which contains the medical information of more than 17,000 AF patients from 32 hospitals in Beijing, China. The time period of CAFR data covers a 5-year collection period from 2011 to 2015. Specifically, the patient dataset contains baseline data, which are the data collected at the time of registry, and the follow-up data. At baseline, the patient information of demographics, symptoms and signs, medical history, results of physical examination and laboratory test, treatments at baseline are collected. Patients are then followed up every 6 months. At each follow-up visit, information about treatments and the clinical events such as AIS are collected. In this paper, different patient groups are selected from the CAFR dataset according to different requirements of the studies, and further details are given in the following subsections.

AIS prediction

The pipeline of building one-year AIS prediction models is illustrated in Figure 1.

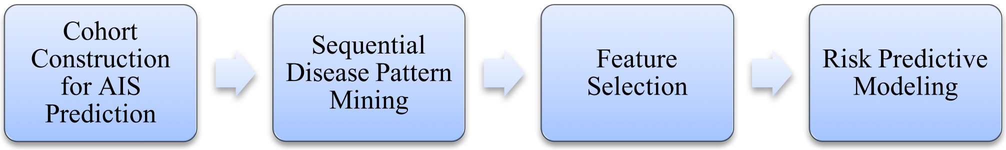

Cohort Construction for AIS Prediction
Sequential Disease Pattern Mining
Feature Selection
Risk Predictive Modeling

Figure 1. Pipeline of the AIS prediction modelling study.

To begin with, the study cohort for AIS prediction is constructed. We select the patients who have not been treated with warfarin or radiofrequency ablation (RFA) at baseline and have completed one-year follow-up records at the same time. A total of 3739 patients who meet our criteria are selected as the patient cohort for the study. Among which, 143 (3.82%) patients are cases whose records show that AIS occurred at least once during the one-year follow-up after registry, and this cohort is highly unbalanced.

The study timeframe covers baseline, the 6th and 12th month follow-ups. Figure 2 shows the timeline and the data information from the selected patient group. Baseline patient record data which are the patient history data of the registry are to be proceed as features of AIS risk models. Baseline data include patient demographics, vital signs, laboratory test results, life styles, treatments, diagnosis history and the corresponding dates of the diagnosis. Among which, diagnosis history and relevant dates will be used to discover the temporal patterns of disease progressions in patient history. The follow-ups in the following year is the AIS prediction period which is used to observe whether an AIS is occurred.

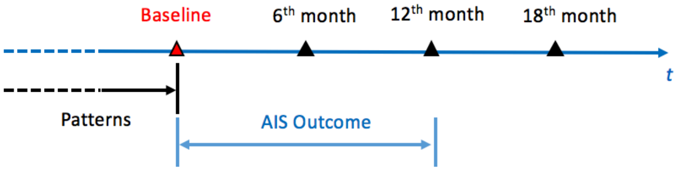


Figure 2. Timeframe for the AIS prediction.

At the second step, sequential pattern mining (SPAM) is performed in order to identify the frequent temporal patterns of patient diagnosis history. SPAM is an algorithm for mining temporal patterns including sequential patterns and combined patterns. Its strategy combines a vertical bitmap representation of the database with efficient support counting[14]. It firstly builds lexicographic trees for generating sequences in different depths from a parent pattern. Then, using Apriori-based pruning technology prunes the candidate extensions of a node in the tree. Finally, it uses bitmap to represent the sequences. In this paper, we apply SPAM as an example of the temporal pattern mining methods, considering the Boolean data types and insufficient time interval data in our dataset, and demonstrate how to integrate it in risk prediction and treatment pathway mining. It is worth noting that other temporal pattern mining methods could also be integrated using our approaches. In this study, all temporal patterns with support >0.01 can be discovered after this step.

Then, the feature selections are performed. The newly discovered sequential patterns are taken as new features. These features combined with original features at single-time points (i.e., baseline time) are selected using relevant feature selection methods. Three main supervised feature selection methods in machine learning are conducted, including:

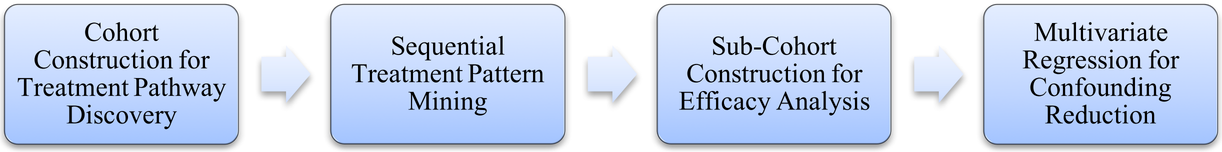
- Filter: In filter feature selections, each feature obtains a score to represent its relevance with the outcome, and the features are ranked according to their scores. In this study, we apply information gain (IG) as the relevance score. IG represents the change in information entropy when a feature is given.
- Wrapper: Wrapper methods evaluate subsets of features in a specific metric to detect the best performance feature subset. In this study, logistic regression (LR) is applied as the learning model and *the area under the receiver operating characteristic curve* (AUC) is taken as the metric. We select the feature subset having the highest AUC value of the LR model by cross validation. The best first search strategy is used in our wrapper selection.
- Embedded optimization: Embedded optimization methods incorporate feature selection into the learning process of a model. In this study, we apply Lasso, *least absolute shrinkage and selection operator*, to perform L1 regularization to LR, which can help to select high relevant features by shrinking the coefficients of low relevant features to zero during model training[17].

After the feature selection is conducted, the potential risk factors are uncovered. Furthermore, we apply logistic regression (LR) to build the one-year AIS risk model based on the selected features. LR is widely used in medical statistics and machine learning research because of its good prediction ability and interpretability.

The risk model is evaluated using randomly repeated cross validation. The model performance is evaluated using AUC which is the standard metric in evaluating prediction models. In addition, the model is also evaluated using *area under the precision recall curve* (AUPR), which helps to provide a more informative assessment for highly unbalanced datasets.

Treatment pathway discovery

The treatment pathway discovery study aims to discover significantly effective treatment pathways for preventing AIS. The study process is illustrated in Figure 3.


Cohort Construction for Treatment Pathway Discovery

Sequential Treatment Pattern Mining

Sub-Cohort Construction for Efficacy Analysis

Multivariate Regression for Confounding Reduction

Figure 3. Pipeline of the treatment pathway study.

During the step of cohort construction, a total of 5635 patients with two-year follow-ups are selected as the objects. Patient data are collected at five time points in a two-stage timeframe, shown in Figure 4. The first time point is the baseline when the patients were registered. Then, patient data are collected every six months after register for two years. Events from the baseline and the first-year follow-ups will be used to identify temporal patterns. The events include the baseline diagnosis, baseline treatments and follow-up treatments. At the same time, the AIS events in the second year of follow-ups are taken as patients' outcomes. According to whether an AIS occurs in the second year, the 5635 patients are divided into a case group of 120 patients and a control group of 5515 patients.

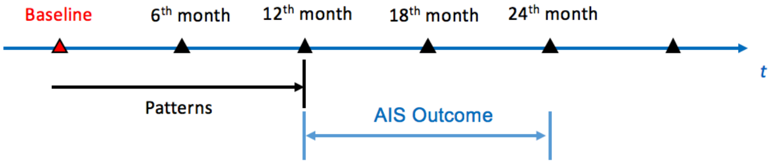


Figure 4. Timeframe for treatment pathways.

The second step is the sequential pattern mining using SPAM. This step discovers all potential treatment pathways which are frequent sequential treatment patterns from the baseline, the 6th and 12th follow-ups. These patterns frequently appear in overall patient group with support >0.05.

However, it has been still unclear whether each pattern is significantly effective for preventing AIS. Therefore, the sub-cohort of the patients with the sequential treatment patterns are constructed, and efficacy analysis is conducted in order to select significantly effective treatment pathways from a large number of patterns.

For each treatment pathway pattern in the SPAM results, we select the patients with the full treatment pathway pattern. In comparison, we also select the patients with its shorter treatment pathway pattern. For example, assuming that a case group of patients have a treatment pattern "medicine A → medicine B", we would select the control group as the patients who take "medicine A" only.

By observing the outcomes of the two groups of patients, we calculate the univariate odds ratios (OR) and p-values for Chi-squared test of each treatment pathway pattern from SPAM results to evaluate the efficacy of the treatment pattern. If the OR of a treatment pattern is less than 1.0 and the p-value is less than 0.1, then this treatment pattern can statistically decrease the risk of AIS occurrence. Furthermore, this study considers the effect of the confounding factors, including CHA2DS2-VASc factors (age, prior TE, congestive heart failure, hypertension, diabetes mellitus, vascular disease and sex), RFA and warfarin. Multivariate logistic regression is implemented to control the confounding factors. The adjusted odds ratios (AOR) and p-values for Wald test are obtained from logistic regression.

The significantly effective treatment pathways with AOR < 1 & p-value <0.1 are finally discovered after controlling the confounding factors.

Results

In this section, we describe the results of the two case studies, one-year AIS prediction and treatment pathway discovery, respectively.

AIS prediction

For AIS risk prediction, the original features are based on single-time points at baseline, of which the number is 53. After applying the SPAM method, a total of 123 frequent sequential disease patterns (support > 0.01) before baseline are found.

In order to observe the effect of the sequential patterns, we compare the prediction performance of the logistic regression model built on the set of original features with that of the model built on the union set of original and sequential features. We evaluate the mean and the standard deviation of AUC and AUPR of each model on 5 randomly repeated 10-fold cross validation partitions of the data, and each model is built based on the same data partitions from the cross validation. It is worth noting that the baseline of AUPR for our dataset is 0.038, which is the average precision of randomly predicting the risk (i.e., an AUPR of 0.076 means the average precision is doubled than that of random prediction). The above modeling and evaluation process is repeated for different feature selection methods.

The results of before-and-after considering sequential patterns are compared in Table 1. When no feature selection is conducted, AUC and AUPR of the prediction models decrease after all the 123 sequential patterns are added to the original feature set. In the filter feature selection experiments, we respectively select top 20 features using the information gain (IG) from both the feature sets. The results show that AUC and AUPR has been improved after considering the temporal features. In the wrapper feature selection experiments, a total of 44 features including 29 temporal features are selected, and the AUC is significantly improved by about 0.3, and the AUPR is also increased by 0.1, compared with the results using original features only. In the embedded feature selection experiments, we implement Lasso with C = 0.5 and select the features whose coefficients in logistic regression are non-zero. The results also appear the improvement in AUC and AUPR after considering the temporal features. Overall, our way of combining sequential pattern mining and feature selection methods improve the performance of AIS prediction.

Table 1. Mean and standard deviation of AUC and AUPR of the LR models built on different feature sets, evaluated by cross validation.

	Original Features			**Original + Temporal features**		
Selection Method	**#OF***	**AUC**	**AUPR**	**#OF, #TF***	**AUC**	**AUPR**
None	53	**0.651±0.008**	**0.063±0.002**	53, 123	0.549±0.017	0.048±0.003
IG (Top 20)	20	0.656±0.005	0.071±0.002	10, 10	**0.671±0.004**	**0.073±0.003**
Wrapper	15	0.708±0.003	0.085±0.003	15, 29	**0.737±0.005**	**0.095±0.005**
Lasso (C = 0.5)	37	0.684±0.005	0.072±0.002	36, 23	**0.691±0.006**	**0.085±0.003**

∗ #OF: the number of selected original features; #TF: the number of temporal features

Furthermore, we compare the performance of traditional models, Framingham[4] and CHA_2DS_2-VASc[5], with the performance of our models. Each model is evaluated by 5 random repetitions of 10-fold cross validation. The AUC and AUPR of the models are compared in Figure 5, which lists the results of the following models: random prediction, Framingham, CHA_2DS_2-VASc, SPAM combined with IG, SPAM combined with wrapper and SPAM combined with LR with Lasso. As we can see, our approaches have better performance in AUC and AUPR than the Framingham and CHA_2DS_2-VASc models.

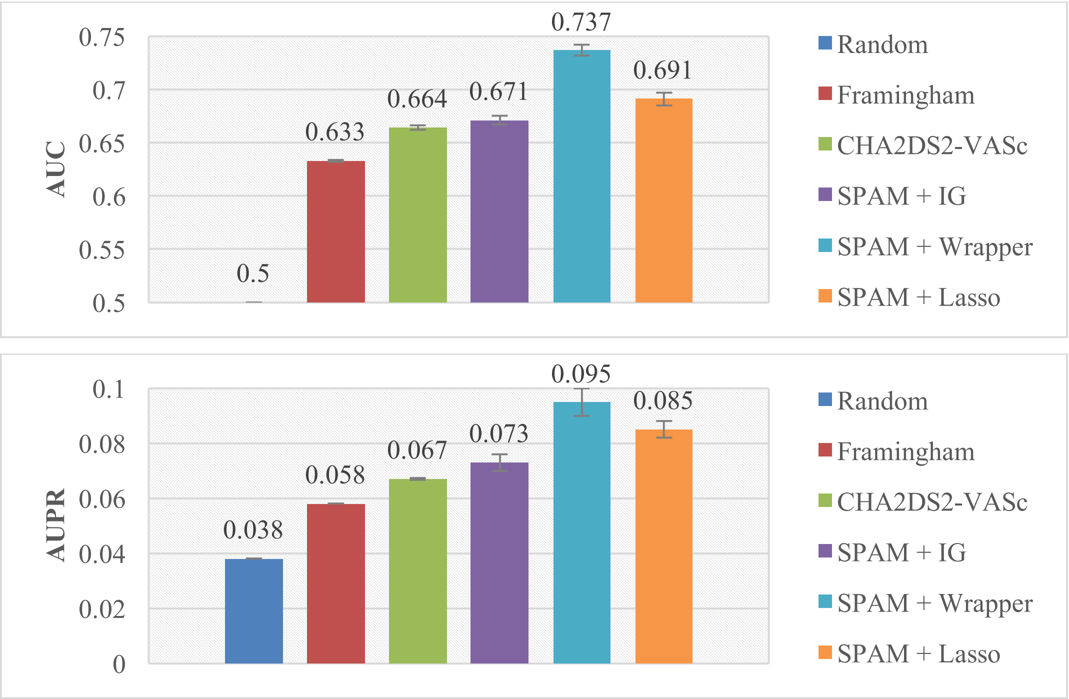


Figure 5. AUC and AUPR of different models by cross validation.

Traditional Framingham score for AIS prediction considers five risk factors which are age, prior stroke/TIA, sex, diabetes mellitus and systolic blood pressure. The risk factors in CHA_2DS_2-VASc models include age, prior TE, congestive heart failure, hypertension, diabetes mellitus, vascular disease and sex. Our approach helps to uncover new potential sequential risk factors for AIS. In Table 2, we list the sequential risk factors which appear more than twice in the above three models: SPAM combined with IG, SPAM combined with wrapper and SPAM combined with LR with Lasso. These discovered risk factors are interpretable and applicable for clinicians. Take the pattern "*Paroxysmal Atrial Fibrillation → Hypertension*" as an example, it suggests that the patients with paroxysmal atrial fibrillation who further develop hypertension could be in a higher risk of AIS.

Table 2. Common Sequential Risk Factors

Common Sequential Risk Factors	
Hypertension → Prior AIS	Hyperlipidemia → Hypertension
Hypertension → Diabetes Mellitus → Heart Failure	Hyperlipidemia → Chronic Atrial Fibrillation
Hypertension → Intracranial Hemorrhage	Hyperlipidemia → Respiratory Disease
AIS → Hyperlipidemia	Diabetes Mellitus → Pacemaker Implantation
Paroxysmal Atrial Fibrillation → Heart Failure	Diabetes Mellitus → Hyperlipidemia
Paroxysmal Atrial Fibrillation → Hypertension	Diabetes Mellitus → Chronic Atrial Fibrillation
Paroxysmal Atrial Fibrillation → Pacemaker Implantation	Chronic Atrial Fibrillation → Prior AIS

Treatment pathway discovery

In this study, we select the patients with the full-length treatments which are the discovered sequential treatments as the case group, and the patients with one less treatments as the control group. 40 statistically significantly effective treatment pathway patterns (support > 0.05, AOR<1, P-value < 0.1) are discovered after controlling the confounding factors. According to the different clinical meaning, these treatment pathway patterns can be classified into two groups: medication continuation, medication efficacy for specific conditions. Table 3 summarizes the two groups and gives relevant result examples with the univariate odds ratios(OR), p-value for Chi-squared test, adjusted odds ratios (AOR) and p-value for Wald test from logistic regression. In the table, the concurrent treatments are presented using "&" , and the sequential treatments are expressed using "→".

Table 3. Treatment pathway pattern results

Pattern Category	Total Number	Pattern Examples				
		Pattern names	**OR**	**P-value for χ^2**	**AOR**	**P-value for Wald**
Medication Continuation	13	Hypertension & Warfarin & ARB* → Warfarin & ARB	0.278	0.075	0.277	0.056
		Hyperlipidemia & Aspirin & Statin → Statin	0.372	0.073	0.346	0.035
Medication Efficacy for Special Conditions	27	Hypertension & ARB → ARB → Warfarin	0.211	0.035	0.224	0.044
		Hyperlipidemia & Aspirin → β-blocker	0.403	0.061	0.392	0.037

∗ ARB: Angiotensin II receptor blocker

The patterns for medication continuation provide the information about the effect of the long-term use of certain medicine from the data. For example, for the pattern "Hypertension & Warfarin & ARB → Warfarin & ARB", the results indicate that the long-term use of warfarin and Angiotensin II receptor blockers (ARB) could be more effective in reducing the risk of AIS for AF patients with the hypertension condition, compared with the discontinuous use of warfarin with ARB. Similarly, continuous use of statin could be more effective compared with the discontinuous use of statin for AF patients who had hyperlipidemia and were taking aspirin.

The patterns for medication efficacy could help to explore the potential effective medicines on specific patient groups. For example, warfarin is significantly effective for the group of patients who also had hypertension and were taking ARB, with no surprise. Similarly, the results also propose that β-blocker is statistically effective for AF patients who also had hyperlipidemia and were taking aspirin.

These treatment pathway patterns are the hypotheses obtained from data evidence, which are statistically significantly effective. This case study shows that our approach could help to propose and explore potential treatment pathway hypotheses based on the data, though relevant clinical trials should be further carried out in order to further prove the efficacy of the findings.

Discussion

The risk prediction results show that the prediction performance will decrease when adding sequential patterns as features, if no feature selection is conducted. The reason is that LR models assume no or little multicollinearity among the features. Adding all sequential patterns can largely increase the correlations among the predictor features, which causes feature redundancy and overfitting problems.

In contrast, our approach of combining SPAM with feature selection methods can largely improve the model performance. In the results, combining SPAM with the wrapper feature selection has the best performance. It is because that wrapper can effectively reduce the feature redundancy. Compared with wrapper, the filter feature selection method selects features according to the relevance scores which are obtained by associating individual features with the outcome, and does not consider the feature correlations. Therefore, it still remains high feature redundancy, and the prediction performance is relatively lower. Besides, Lasso aims to select high variance features, such as numeric values of laboratory tests. However, the sequential patterns in this study are binary features, which have relatively low variance. Thus, the results in combining SPAM and LR with Lasso appear no significant improvement in AUC.

In terms of the temporal pattern mining methods, the SPAM approach is adopted as an example of temporal pattern mining methods in this paper. SPAM can be applied to discovery the sequential patterns only, and the patterns in its findings do not cover other information such as time intervals or knowledge-based abstraction. Nevertheless, several other temporal pattern mining methods could also be selected and implemented in our integration approach, such as KBTA or KarmaLego[15], which can help to discovery more complex temporal patterns for risk prediction and treatment pathway mining.

During the process of treatment efficacy analysis, we apply multivariate logistic regression and obtain the adjusted odds ratios in order to control the confounding factors. Still, we can further implement other confounding reduction methods in the future. For example, we can use propensity score matching to resample the case and control groups. In addition, these treatment hypotheses and the risk factors in the prediction models could be compared with the existing clinical knowledge in the future work, and should be further verified using specific clinical trials.

Also, the additional diagnoses made in the time intervals are not considered in this study because of the uncompleted diagnostic information in the follow-ups.

Furthermore, the approach of this paper could be time consuming when applied in big datasets. It is because the SPAM algorithm extracts all possible frequent patterns at first, and it is time consuming to conduct feature selections or efficacy analysis for each pattern to select the desired patterns from a large number of patterns. Thus, further work could be integrating the feature selections and the efficacy analysis into the process of generating sequential patterns, and undesired patterns can be pruned during the process.

Conclusion

This paper demonstrates the way of integrating temporal pattern mining into AIS risk prediction and treatment pathway discovery for AF patients. We combine temporal pattern mining with feature selection to identify temporal risk factors that have predictive ability, and integrate temporal pattern mining with treatment efficacy analysis to discover significant effective treatment pathway patterns. Results show that our approach has discovered the new potential temporal risk factors for AIS, and has significantly improved AUC and AUPR of the one-year AIS prediction model. Also, our approach has discovered treatment pathway patterns that are statistically effective for AIS prevention of AF patients.

References

1. January CT, Wann LS, Alpert JS, Calkins H, Cigarroa JE, Cleveland JC Jr, et al. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation. Journal of the American College of Cardiology. 2014;64(21):2246–80.
2. Du X, Ma C, Wu J, Li S, Ning M, Tang R, et al. Rationale and design of the Chinese Atrial Fibrillation Registry Study. BMC Cardiovascular Disorders. 2016 Jun 7;16:130. doi: 10.1186/s12872-016-0308-1.
3. Go AS, Mozaffarian D, Roger VL, Benjamin EJ, Berry JD, Blaha MJ, et al. Heart disease and stroke statistics–2014 update: a report from the American Heart Association. Circulation. 2014;129:e28-e292.
4. Wang TJ, Massaro JM, Levy D, Vasan RS, Wolf PA, D'Agostino RB, et al. A risk score for predicting stroke or death in individuals with new-onset atrial fibrillation in the community: the Framingham Heart Study. Journal of the American Medical Association. 2003; 290 (8): 1049-1056.
5. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest. 2010;137:263-72.
6. Li X, Liu H, Du X, Zhang P, Hu G, Xie G, et al. Integrated Machine Learning Approaches for Predicting Ischemic Stroke and Thromboembolism in Atrial Fibrillation. American Medical Informatics Association, 2016
7. Li X, Liu H, Du X, Hu G, Xie G, Zhang P. Using Frequent Item Set Mining and Feature Selection Methods to Identify Interacted Risk Factors - The Atrial Fibrillation Case Study. Studies in Health Technology and Informatics. 2016;228:562-6.
8. Verheugt FW, Granger CB. Oral anticoagulants for stroke prevention in atrial fibrillation: current status, special situations, and unmet needs. Lancet. 2015;386:303-10.
9. Moskovitch R and Shahar Y. Medical temporal knowledge discovery via temporal abstraction. American Medical Informatics Association, 2009.

10. Wang F, Liu C, Wang Y, Hu J, and Yu G. A Graph Based Methodology for Temporal Signature Identification from EHR. American Medical Informatics Association, 2015.
11. Srikant R and Agrawal R. Mining sequential patterns: generalizations and performance improvements. In EDBT, 1996.
12. Zaki MJ. Spade: An efficient algorithm for mining frequent sequences. Machine learning, 42(1-2): 31–60, 2001
13. Han J, Pei J, Mortazavi-Asl B, Pinto H, Chen Q, Dayal U, and Hsu MC. PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. In ICDE, 2001.
14. Ayres J, Flannick J, Gehrke J, and Yiu T. Sequential pattern mining using a bitmap representation. In SIGKDD, 2002.
15. Shahar Y and Musen MA. Knowledge-based temporal abstraction in clinical domains. Artificial intelligence in medicine, 8(3):267–298, 1996.
16. Liu C, Zhang K, Xiong H, Jiang G, Yang Q. Temporal skeletonization on sequential data: patterns, categorization, and visualization. IEEE Transactions on Knowledge and Data Engineering 28 (1), 211 – 223
17. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B. 1996: 267–88.