Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210167

Prediction of In-Hospital Mortality from Administrative Data: A Sequential Pattern Mining Approach

Jessica PINAIRE[a,b,1], Etienne CHABERT[b], Jérôme AZÉ[b], Sandra BRINGAY[b,c], Pascal PONCELET[b] and Paul LANDAIS[a]
[a] *Montpellier university UPRES EA 2415, Clinical Research University Institute*
[b] *Montpellier University LIRMM, UMR 5506,*
[c] *Paul Valery University, AMIS, Montpellier, France*

Abstract. Study of trajectory of care is attractive for predicting medical outcome. Models based on machine learning (ML) techniques have proven their efficiency for sequence prediction modeling compared to other models. Introducing pattern mining techniques contributed to reduce model complexity. In this respect, we explored methods for medical events' prediction based on the extraction of sets of relevant event sequences of a national hospital discharge database. It is illustrated to predict the risk of in-hospital mortality in acute coronary syndrome (ACS). We mined sequential patterns from the French Hospital Discharge Database. We compared several predictive models using a text string distance to measure the similarity between patients' patterns of care. We computed combinations of similarity measurements and ML models commonly used. A Support Vector Machine model coupled with edit-based distance appeared as the most effective model. Indeed discrimination ranged from 0.71 to 0.99, together with a good overall accuracy. Thus**,** sequential patterns mining appear motivating for event prediction in medical settings as described here for ACS.

Keywords. Hospital discharge database, Data mining, In-hospital mortality, Prediction, Machine Learning, Medical event history, Sequential pattern, Acute coronary syndrome.

1. Introduction

Sequence prediction have many application domains such as web page prefetching, product recommendation, stock market prediction, weather forecasting or sequence prediction of clinical events. Consequently, various models have been developed based either on machine learning methods, Markov models, directed graphs, or neural networks models [1–3], grammar inference [4] or process mining [5]. Review of the literature showed that methods based on machine learning techniques outperform other models. To reduce model complexity, a number of solutions have been proposed including combination of pattern mining techniques with pattern matching techniques [2]. On the basis of these findings, we explored such techniques for sequence prediction.

In a previous work, we highlighted the interest of patients trajectories as a decision tool [6]. Our present objective is to show that this tool can be useful for predicting

[1] Corresponding Author, Dr J Pinaire LIRMM, UMR 5506, 860 rue de Saint Priest – Bât 5, 34095 Montpellier, Cedex 5, France. E-mail: jessica.pinaire@lirmm.fr

hospital mortality. It was applied to in-hospital mortality linked to acute coronary syndrome (ACS).

2. Material and methods

The ACS dataset was collected from the French Hospital Discharge Database for the 2009-2014 period [7]. Discharge summaries were extracted according to the International Classification of Diseases 10th revision (ICD-10) codes: I21 to I24 together with the percutaneous coronary intervention codes. A previous work presented the database [8]. Data were de-identified. This study was approved by the Commission Nationale de l'Informatique et des Libertés, agreement No. 1375062.

We included 4,871 patients French metropolitan population >45 years old who experienced at least 4 stays related to cardiovascular diseases, in whom 668 in-hospital deaths occurred. For each discharge summary, a sequence of ICD-10 codes were identified, and called "*patient trajectories*". The dataset was divided into "*contexts*", according to sex, age and number of hospitalizations [8]. Two classes of age have been defined: 45-65 (45% of the sample) and > 65 (55% of the sample). The average number of hospitalizations was 5. The data flow chart appear on Figure 1.

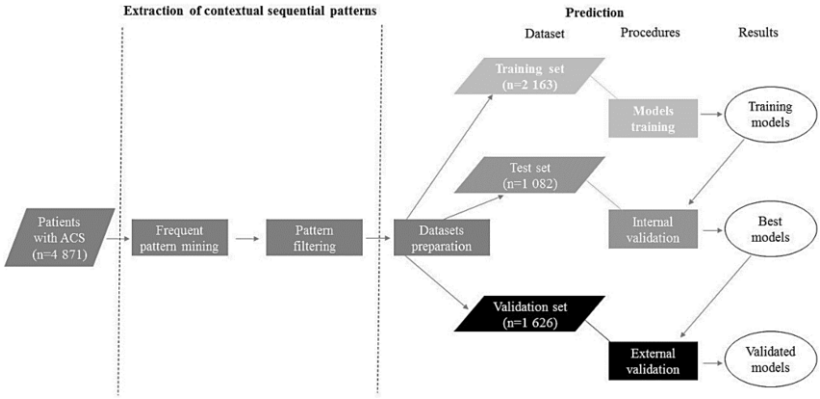

Figure 1. Data flow chart.

Extraction of contextual sequential patterns: Each patient has a list of time-ordered events corresponding to ICD-10 codes (R07: Pain in throat and chest; I20: Angina pectoris; I21: Acute myocardial infarction (AMI); I25: Chronic ischemic heart disease; I50: Heart Failure). An *itemset* it_i is a non-ordered group of events (*items)*, occurring at the same time. A *sequence* $S = \langle it_a, it_b, \dots, it_p \rangle$ is a non-empty and ordered list of p *itemsets*. Sequences are associated with a context. We mined contextual frequent patterns in ACS trajectories using the Contextual Frequent Pattern Mining (CFPM) [9], based on the *PrefixSpan* algorithm [10]. To avoid redundant information increasing the risk of collinearity in predictive models, we filtered the results, retaining the maximal frequent patterns [18], and obtained a list of maximum frequent patterns by context [11]. We determined the best modeling and associated predictive performance. The objective was to predict mortality occurring in a care facility. A 4-step procedure was designed according to the TRIPOD guidelines [12].

Dataset preparation: the dataset randomly split into 2 samples. The first one (n=3245) comprised two sub-samples: one for models training (n = 2163), and the other

for internal validation (n = 1082). The second sample (n = 1626) was retained for external validation. We integrated the patterns discovered in the previous module as predictors, by measuring the *similarity* between these patterns and patients' trajectory. The similarity measure was integrated in the choice of the model. Similarities were calculated for the following distances: longest common substring distance, Levenshtein distance, optimal string alignment distance, Damerau-Levenshtein distance, q-gram distance, Jaccard distance, cosine distance, Jaro distance and Jaro-Winckler distance.

Training models: Predictors were sex, age group and similarities continuous or discretized. Based on a cross-validation principle with training and test sets, we compared most popular models: Naïve Bayes (*NB*), k-nearest neighbours algorithm (*KNN*), Regression tree (*Tree*), Logistic regression (*LR*), Support Vector Machine (*SVM*) and Artificial Neural Networks (*ANN*).

Internal validation. We assessed the quality of the prediction by calculating the discrimination with the following criteria: accuracy, sensibility, specificity, error rate, precision, F-measure, Area Under ROC Curve (AURC). Based on these discrimination measures, we chose the *(model, similarity)* combination presenting the best compromise using the maximal vector computation method.

External validation. We evaluated the discrimination power and overall accuracy of selected models. Discrimination was assessed by AURC and accuracy by Brier score.

3. Results

The prediction module evaluated 6 models, 9 similarity measures and 2 types of variables (discretized and continuous) i.e. 108 different models. Best combinations of *(model, similarity)* were explored. Internal validation used the key metrics grouped by categories of models and similarities in every contexts. Table 1 presents the best combinations (in bold) resulting from the selection process. In most contexts the best combinations were (SVM, edition). In addition, models with continuous similarities performed better than those with discretized similarities. We found eight combinations with heuristic similarities associated with SVM, ANN and LR models, essentially in the ≤5 stays group contexts. We also found six combinations with q-gram similarities that performed better.

Then, we aggregated the results by context, for each type of model, and ranked the models according to their performance. This ranking in percentage for the three best performances appear on table 2. For ICD-10 codes trajectory modeling, SVM was the most efficient model in 74% of cases, then ANN models (70%) and LR models (43%).

Table 1. Distribution (%) of the best combinations (model, similarity) according to ICD 10 codes trajectories.

	ICD-10 code trajectories			
	Tree	**LR**	**SVM**	**ANN**
edition	-	2.86	**42.86**	17.14
q-gram	-	5.71	5.71	2.86
heuristic	5.71	-	11.43	5.71

We explored the performances of the models. AURC ranged from 0.71 to 0.99 for ICD-10. According to this criterion, the best results were found in the following contexts: Women & 45-65 years, ≤5 stays and also Men & 45-65 years & ≤5 stays. Conversely, the worst models concerned the >65 years & > 5 stays.

Table 2. Average ranking (%) of the best models across all contexts and similarities.

Rank	NB	KNN	Tree	LR	SVM	ANN
1st	-	-	-	4.35	**73.91**	21.74
2nd	4.35	-	4.35	4.35	17.39	**69.57**
3rd	30.43	-	13.04	**43.48**	8.70	4.35

External validation: In the final step, we proceeded to an external validation. AURC varied from 0.57 to 0.91 for ICD-10. The higher AURC values were found in 45-65 years & ≤5 stays. Less discriminant case concerned the context of 45-65 years & >5 stays. In parallel, Brier score ranged from 0.09 to 0.26. The best overall accuracy was found in >65 years & ≤5 stays. In contrast, the worst accuracy appeared for the context of 45-65 years & >5 stays.

4. Discussion

We used sequential patterns to elaborate in-hospital mortality prognostic models. Sequential patterns were integrated as predictors by measuring a similarity between patients' trajectory and patterns. We compared the most popular string distances. We used the most commonly predictive models for comparison. Our purpose was to establish the best (model, similarity) combination by context. The originality of our work is to use patient trajectories as predictors through similarity scores, while considering the medical singularity of each type of population.

In most contexts, SVM model coupled with edit-based distance was the most efficient combination associated with in-hospital mortality. In most of cases, ANN were the second more effective models, followed by LR models. All three shared quite equivalent performances in terms of calibration and discrimination. Of note, LR models provided satisfactory results in predicting in-hospital mortality in patients with Acute Myocardial Infarction. In addition, comparing ANN, SVM and LR models for mortality prediction in patients with cardiovascular diseases, differences were not significant between machine learning models and classical regression models [13]. In another work, decision trees outperformed LR, ANN and SVM algorithms in mortality prediction, but for intensive care unit data. Furthermore, a review of risk prediction models for electronic health records data, reported that linear regression models were the most common algorithms used with a high level of accuracy [14].

Model combinations with edit-based distance were often the most efficient. For string distances, the choice usually depends on the nature of the data and the length of the sequences. Thus, q-gram distances appeared well suited for very long length sequences, contrarily to heuristic distances [14]. Thus, we observed a q-gram distance associated with the model essentially for contexts including the >5 stays category and/or the >65 years age group for whom the length of trajectories were substantially longer. Conversely, a heuristic distance appeared more frequently for contexts including ≤5 stays category and/or the 45-65 years category for which trajectories were potentially shorter associated with younger age. Thus, our results were consistent with the choice of the final distance selection as well as the length of the sequences.

A comparison study evaluating established risk prediction models for cardiovascular disease showed that performances varied from 0.71 to 0.88 according to the AURC criterion close to our results ranging from 0.71 to 0.99.

Our work has some limitations. We did not create different contexts by using either comorbidities or the type of care procedure to mine patterns more specific to a sub-population; neither we combined DRG and ICD-10 code sequences, or even added information such as related diagnoses, care procedures or comorbidity scores. Sequences of several *itemsets* could have been introduced instead of one. For the prediction module other models might have been proposed: random forest, boosted or regression trees; features selection with techniques like wrappers, filters or embedded methods; or tuning of the final models while adjusting their parameters with optimization algorithms.

As a conclusion, sequential patterns mining appear motivating for event prediction in medical settings as described here for ACS, with several applications in medical practice. Thus, as a monitoring tool, it might contribute to measure the burden of disease and improve healthcare. Moreover, a risk score might be useful for patients' triage and provide a decision-support tool to help orienting towards the most convenient care strategy. In a public health perspective, a better knowledge of the relationship between care pathways, comorbidities and mortality might be an aid to medical decision making.

References

[1] Laird P, Saul R. Discrete sequence prediction and its applications. Mach Learn. 1994;15:43–68.
[2] Pitkow J, Pirolli P. Mining longest repeating subsequences to predict world wide web surfing. In: Proceedings of USITS' 99: The 2nd USENIX Symposium on Internet Technologies & Systems. Boulder, Colorado, USA; 1999. p. 1–13.
[3] Gueniche T, Fournier-Viger P, Tseng VS. Compact Prediction Tree: A Lossless Model for Accurate Sequence Prediction. In: Motoda H, Wu Z, Cao L, Zaiane O, Yao M, Wang W, editors. Advanced Data Mining and Applications. Springer; 2013. p. 177–88..
[4] Datta S, et al. A Grammar Inference Approach for Predicting Kinase Specific Phosphorylation Sites. PloS One. 2015;10:e0122294.
[5] Van der Aalst WMP, et al. Time prediction based on process mining. Inf Syst. 2011;36:450–75.
[6] Pinaire J, Azé J, Bringay S, Landais P. Patient healthcare trajectory. An essential monitoring tool: a systematic review. Health Inf Sci Syst. 2017;5:1–18.
[7] Chantry AA, Deneux-Tharaux C, et al. Hospital discharge data can be used for monitoring procedures and intensive care related to severe maternal morbidity. J Clin Epidemiol. 2011;64:1014–22.
[8] Pinaire J, Azé J, Bringay S. ,net al. Hospital Burden of Coronary Artery Disease: Trends of Myocardial Infarction and/or Percutaneous Coronary Interventions in France 2009-2014. PloS One. 2019;14:e0215649.
[9] Rabatel J, Bringay S, Poncelet P. Mining Sequential Patterns: A Context-Aware Approach. In: Guillet F, Pinaud B, Venturini G, Zighed DA, editors. Advances in Knowledge Discovery and Management. Springer Berlin Heidelberg; 2013. p. 23–41.
[10] Han J, Pei J, Mortazavi-Asl B, Pinto H, Chen Q, Dayal U, et al. PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. In: Proceedings of the 17th international conference on data engineering. 2001. p. 215–224.
[11] Yang G. The Complexity of Mining Maximal Frequent Itemsets and Maximal Frequent Patterns. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2004. p. 344–353.
[12] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.
[13] Christodoulou E, Ma J, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22.
[14] Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. 2017;24:198–208.