Article

Patterns

Transitive Sequencing Medical Records for Mining Predictive and Interpretable Temporal Representations

Highlights

- A transitive sequential pattern mining (TSPM) approach is presented for clinical data
- TSPM mines temporal sequences from discrete electronic health records (EHR) data
- TSPM representations from EHRs are superior disease "differentiators" and predictors
- An ML pipeline is introduced to apply TSPM in phenotype prediction and classification

Authors

Hossein Estiri, Zachary H. Strasser, Jeffery G. Klann, ..., Victor M. Castro, MaryKate E. Murphy, Shawn N. Murphy

Correspondence

hestiri@mgh.harvard.edu

In Brief

Electronic medical records (EHRs) contain valuable temporal information about progression of diseases and treatment trajectories. However, time is not precisely captured in clinical data. In a study of congestive heart failure, Estiri and colleagues propose an approach for extracting temporal sequential patterns from EHR data that improve phenotype prediction and classification and are also interpretable.

Estiri et al., 2020, Patterns *1*, 100051
July 10, 2020
https://doi.org/10.1016/j.patter.2020.100051

CellPress

Please cite this article in press as: Estiri et al., Transitive Sequencing Medical Records for Mining Predictive and Interpretable Temporal Representations, Patterns (2020), https://doi.org/10.1016/j.patter.2020.100051

Patterns

CellPress
OPEN ACCESS

Article

Transitive Sequencing Medical Records for Mining Predictive and Interpretable Temporal Representations

Hossein Estiri,[1,2,3,8,*] Zachary H. Strasser,[1,2,3,4] Jeffery G. Klann,[1,2,3] Thomas H. McCoy, Jr.,[3,6] Kavishwar B. Wagholikar,[1,2,3] Sebastien Vasey,[5] Victor M. Castro,[2] MaryKate E. Murphy,[2] and Shawn N. Murphy[1,2,3,4,7]

[1]Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA 02144, USA
[2]Research Information Science and Computing, Mass General Brigham, Somerville, MA 02145, USA
[3]Harvard Medical School, Boston, MA 02115, USA
[4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
[5]Department of Mathematics, Harvard University, Cambridge, MA 02138, USA
[6]Center for Quantitative Health, Massachusetts General Hospital, Boston, MA 02114, USA
[7]Department of Neurology, Massachusetts General Hospital, Boston, MA 02114, USA
[8]Lead Contact
*Correspondence: hestiri@mgh.harvard.edu
https://doi.org/10.1016/j.patter.2020.100051

THE BIGGER PICTURE Over the past decade, billions of dollars have been spent to institute meaningful use of electronic health record (EHR) systems. For a multitude of reasons, however, EHR data are still complex and have ample quality issues, which make it difficult to leverage these data to address pressing health issues, especially during pandemics such as COVID-19, when rapid responses are needed. In this paper, we propose a transitive sequential pattern mining algorithm for exploiting the temporal information in the EHRs that are distorted by layers of administrative and healthcare system processes. Perhaps more importantly, we propose a machine learning (ML) pipeline that is capable of engineering predictive features without the need for expert involvement to model diseases and health outcomes. Together, the temporal sequences and the ML pipeline can be rapidly deployed to develop computational models for identifying and validating novel disease markers and advancing medical knowledge discovery.

Proof-of-Concept: Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

Electronic health records (EHRs) contain important temporal information about the progression of disease and treatment outcomes. This paper proposes a transitive sequencing approach for constructing temporal representations from EHR observations for downstream machine learning. Using clinical data from a cohort of patients with congestive heart failure, we mined temporal representations by transitive sequencing of EHR medication and diagnosis records for classification and prediction tasks. We compared the classification and prediction performances of the transitive sequential representations (bag-of-sequences approach) with the conventional approach of using aggregated vectors of EHR data (aggregated vector representation) across different classifiers. We found that the transitive sequential representations are better phenotype "differentiators" and predictors than the "atemporal" EHR records. Our results also demonstrated that data representations obtained from transitive sequencing of EHR observations can present novel insights about the progression of the disease that are difficult to discern when clinical data are treated independently of the patient's history.

Please cite this article in press as: Estiri et al., Transitive Sequencing Medical Records for Mining Predictive and Interpretable Temporal Representations, Patterns (2020), https://doi.org/10.1016/j.patter.2020.100051

INTRODUCTION

The widespread adoption of electronic health records (EHRs) has accumulated an unprecedented amount of patient health information. EHRs contain important temporal information that provide an opportunity for discovering new insights about disease progression and treatment trajectories. However, EHR observations reflect a complex set of processes that thwart their seamless translation into actionable knowledge. Namely, the raw EHR records may not be direct indicators of patients' "true" health states at different time points, but rather reflect the clinical processes (e.g., policies and workflows of the provider and payor), the patients' interactions with the system, and the recording processes.[1–3]

Biomedical researchers increasingly apply conventional association studies to EHR data, yet the temporality of these data have not been fully exploited by current methods.[1,4] The temporal properties of EHR data signify the complexities of the healthcare processes. This makes incorporating temporal information into common temporal analysis techniques challenging.[5] The challenge is twofold. First, EHR observations are recorded asynchronously across time (i.e., measured at different times and irregularly), which provide foundational challenges for directly applying common temporal analysis methods.[6–9] Second, translating the temporal nature of discrete EHR data into useful data representations (or features) for standard machine learning (ML) algorithms is challenging.[10,11] This is a critical gap. This study aims to address this gap by developing and testing a temporal representation mining algorithm for asynchronously recorded discrete EHR data.

We propose a transitive sequencing algorithm for constructing temporal representations from medications and diagnoses data from EHR. We conduct this research with the application in classifying and predicting congestive heart failure (CHF) in patients. Our results demonstrate that temporal sequences of electronic medical records improve computational classification of patient cohorts, and phenotype prediction, when no record of the phenotype exists in the medical records. The proposed transitive sequential representations are interpretable and also more predictive features for standard ML algorithms than "atemporal" representations of discrete EHR data. We found that the sequential representations improve CHF classification by over 4% and prediction by over 13%. We also demonstrate that the proposed transitive sequential representations are more suited than the sequences mined through traditional sequential pattern mining (SPM) algorithms for ML using clinical data that inherit various biases due to the recording process.

Background

An extended body of work exists on extracting interval-based symbolic representations from clinical measurement data (e.g., laboratory test results),[12,13] often known as the temporal abstraction (TA) approach.[14,15] Although forms of such representations have been utilized as features in classification/prediction tasks,[16–19] application in ML is not the focus in the TA agenda. Furthermore, the development of TA methods has been largely confined to continuous clinical measurements data.[12,20,21] In addition to continuous data, however, EHRs contain discrete data, such as records of diagnoses, medications, and procedures.

For discrete clinical data, SPM[22] approaches, such as the frequent sequential pattern (FSP),[22] are promising. The goal in SPM is to discover relevant subsequences from a large set of sequences (of events or items) with time constraints. Several SPM algorithms have been developed to improve mining efficiency and address specific domain needs (for recent surveys of SPM algorithms, refer to the studies by Fournier-Viger et al.[23,24] and Truong-Chi and Fournier-Viger[23,24]). As a result, SPM algorithms are fairly mature in computational and data management schemes. However, as the SPM approaches were primarily developed for transaction data, the importance of a sequential pattern for use in downstream ML algorithms is often determined by an occurrence frequency threshold, known as the minimum support.[25,26] The goal is to cut the number of data representations by finding the most frequent temporal patterns among all patterns. This strategy has been used by the majority of the literature using SPM approaches for clinical data mining.[4,7,10,26–29] However, because the temporal patterns are mined based on frequency, some may not make clinical sense or do not apply to clinical data that inherit various biases through the recording process.

A priori-based SPM methods are popular in the healthcare domain. The a priori property is that if a sequence cannot pass the minimum support test (i.e., is not assumed frequent), all of its subsequences will be ignored. Examples of a priori-based algorithms with use cases in clinical data are SPM with regular expression constraints (SPIRIT),[30] sequential pattern discovery using equivalence classes (SPADE),[31] and SPM (SPAM).[32]

In particular, a couple of studies have applied adjustments to the SPM's frequency-based criterion for feature selection using clinical data. Liu et al.[33] proposed a temporal graph approach to predict the onset of CHF by summarizing multiple sequences of events recorded for a patient.[33] Due to added complexities in the network of events encapsulated in the temporal graphs, the authors had to work through a specialized generalization method. Guo et al.[34] showed the efficiency of sequential patterns in predicting the risk of acute ischemic stroke over Framingham and CHA_2DS_2-VASc models.[34] They applied feature selection procedures to reduce the dimensionality of the temporal patterns mined by the popular SPM algorithm (SPAM[32]).

Considering EHRs as "indirect" reflections of a patient's true health state, we propose an algorithm for mining transitive sequential patterns from discrete EHR data, apply dimensionality reduction, and implement the top features in phenotype classification and prediction. Our approach provides a specification for SPM and a formal dimensionality procedure—minimize sparsity and maximize relevance (MSMR)—that integrates feature selection into the classification task.

RESULTS

Study design is illustrated in Figure 1. A summary of the patient characteristics is provided in Table 1. Changes in the number of unique medication and diagnosis records through feature engineering and initial dimensionality reduction processes are presented in Table 2. For classification, we began with more than 45,000 unique medication/diagnosis records in the aggregated vector representation (AVR) approach, from which we mined over 137 million unique transitive sequence representations.

Please cite this article in press as: Estiri et al., Transitive Sequencing Medical Records for Mining Predictive and Interpretable Temporal Representations, Patterns (2020), https://doi.org/10.1016/j.patter.2020.100051

CellPress
OPEN ACCESS

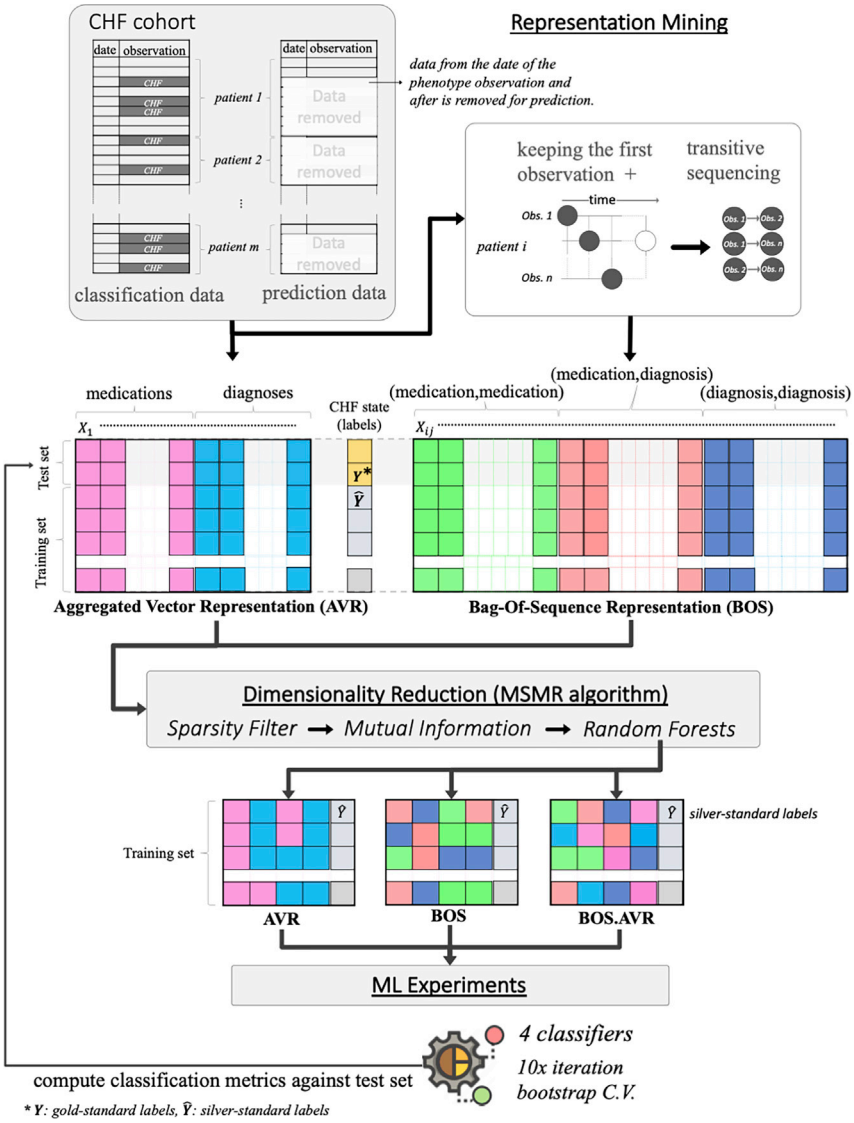

Figure 1. The Study Design Encompasses Representation Mining, Dimensionality Reduction, and ML Experiments

These numbers were over 25,000 records and 30 million sequenced representations in the prediction task. Sparsity screening at the threshold of 1% (68 patients for classification and 58 patients for prediction) resulted in 6,469 AVR representations (i.e., unique medication/diagnosis records) and over 1,300,000 sequenced representations in the classification task. For prediction, the sparsity screening resulted in a reduction in dimensionality to over 2,500 AVR and 100,000 sequenced representations.

At the conclusion of the MSMR algorithm, for each approach the top 200 representations were selected based on their mean decrease in Gini (MDG) (in tied situations, mutual information and prevalence were used). In the next section we compare the classification results obtained from the four different classifiers using the top 200 features.

In general, we found that the bag-of-sequences (BOS) approach outperforms the AVR approach for both classification and prediction tasks. We review the results by hypothesis. Testing median and best area under the receiver-operating characteristic (ROC) curve (AUC-ROC) values from each approach, classifier, and feature set are provided in Table 3, and performance metrics from cross-validation are reported in Table S1. Figure 2 presents the classification (top plot) and prediction (bottom plot) AUC values as well as the distribution of those values across all three approaches (AVR, BOS, and BOS-AVR), and by classifiers.

Classification

We found that, across the four classifiers, the BOS representations provided an improvement in classification performance over the AVR representations by an average of 4.1% AUC-ROC (Table 3). The average median classification AUC-ROC values were 0.81 (AVR), 0.84 (BOS), and 0.85 (BOS-AVR). The best overall classification performances were from model averaged neural networks (MA Neural Net.) that resulted in AUC-ROC values of 0.87, 0.88, and 0.92 using the AVR, BOS, and the BOS-AVR data representations, respectively. For the purpose of classification, combining data representations from the AVR and BOS approaches resulted in the best performance (AUC-ROC = 0.92).

Prediction

For predicting heart failure, the performance improvements provided by the BOS representations were even more substantial when using medication and diagnosis records from prior to the first diagnosis of the phenotype in the medical records. On average, the BOS representations improved median prediction performance from AVR representations by 13% in AUC-ROC (Table 3). The average median prediction AUC-ROC values were 0.72 (AVR), 0.81 (BOS), and 0.72 (BOS-AVR). For the purpose of prediction, the sequenced data representations (BOS) provided the best performances (AUC-ROC = 0.87). When the AVR and BOS representations were combined, the prediction performance was inferior to BOS-only representations. Using the data from before the first record of the CHF, the best overall prediction performances were from support vector machines with class weights (SVM CW) with AUC-ROC values of 0.82, 0.87, and 0.83, respectively from the AVR, BOS, and the BOS-AVR data representations.

Sequential Pattern Mining

As described in Experimental Procedures, we also mined the traditional sequential patterns and selected the most frequent to represent the SPM for comparison. We only compared the

Please cite this article in press as: Estiri et al., Transitive Sequencing Medical Records for Mining Predictive and Interpretable Temporal Representations, Patterns (2020), https://doi.org/10.1016/j.patter.2020.100051

Table 1. Summary of Patient Characteristics in the Test and Training Sets

	Test Set (N = 56)	Training Set (N = 6,851)
Gender	female: 52%	female: 42%
	male: 48%	male: 58%
Race	white: 86%	white: 86%
	black: 9%	black: 7%
	Asian: ~0%	Asian: 1%
	other/unknown: 5%	other/unknown: 6%
Ethnicity	Hispanic: ~0%	Hispanic: 4%
Observation range	mean: 17 years (SD: 6.5)	mean: 16 years (SD: 7.8)
Age	mean: 72 years (SD: 14.7)	mean: 68 years (SD: 13.8)

results of the SPM against BOS with regularized logistic regressions as this was not the focus of this study. As illustrated in Figure 2, in the classification task, the SPM approach's performance was only slightly inferior to that of BOS, but statistically inferior to the BOS-AVR approach. Similar to the BOS and the BOS-AVR approaches, the SPM resulted in better classification performance than the AVR. However, the transitive sequencing algorithm demonstrated an unparalleled improvement over the SPM in prediction: median AUC-ROC 0.634 (SPM) versus 0.79 (BOS). Given these results, one can conclude that the transitive sequencing algorithm is more suited for modeling clinical data in comparison with the conventional SPM.

Clinical Interpretations

We used the visual dashboard to further explore the top 200 sequences for classification and prediction. A snapshot of the visualizations and functionalities of the dashboard is provided in Figure 3. The landing page in the dashboard provides an interactive flow diagram (Sankey plots) constructed from the classification/prediction sequences identified by the MSMR algorithm. The user has the ability to zoom into specific sequences by either selecting the sequence or specifying the rank threshold. In addition, the dashboard provides queryable tabular pages that present metrics including the prevalence, mutual information, and MDGs for selected sequences. Additionally, it provides donut chart visualizations of the likelihood of heart failure given a specific sequence versus the CHF likelihood for the individual elements of the sequence (for examples see Figure 4).

The transitive sequences are, in general, better "differentiators" for identifying heart failure than the "atemporal" EHRs. The signal obtained from individual features as to whether a patient truly has (or does not have) heart failure often strengthens when the features are in sequences. In Figure 4, the diagnoses codes for "heart failure," "chronic obstructive pulmonary disease," and "benzodiazepines" give probabilities of CHF at 45%, 47%, and 63%, respectively. However, when these features are sequenced with one another, the probability of heart failure increases. For example, the sequence "heart failure → benzodiazepine" has a likelihood of 64% for heart failure and "heart failure → other chronic obstructive pulmonary disease" has a likelihood of 78%. The temporal sequences confer a greater signal that a patient truly has heart failure compared with the raw elements (i.e., AVR features).

From among the top sequences, clinical experts identified those sequences that match a common clinical narrative among patients with heart failure versus those who lack an obvious clinical explanation. Table 4 has specific examples of the two groups of transitive sequences. For example, the sequence "abnormalities in breathing → cardiomyopathy" matches a common clinical scenario in heart failure patients. One would expect a heart failure patient to have the symptom of difficulty breathing, and it is likely for the patient to then be given the diagnosis of cardiomyopathy based on subsequent imaging studies. Another sequence that is easy to interpret is "heart failure → metoprolol." Metoprolol is a common medication for heart failure and is frequently started after a diagnosis of heart failure. A less obvious sequence is "topical anti-infectives → unspecified kidney failure." Neither of these two components are clearly related to heart failure in the same way that difficult breathing, cardiomyopathy, and metoprolol are related to heart failure.

DISCUSSION

Using transitive sequences of EHR observations, we constructed data representations that are both predictive and interpretable. In the context of phenotyping CHF patients, our results demonstrate that harnessing the knowledge of disease progression through temporal sequencing (the BOS approach) improves classification and prediction over the conventional approach (AVR). The classification and prediction performances obtained in this study are comparable with the state-of-the-art classification/prediction models for heart failure. For example, the highest median AUC-ROC in Wu et al.[35] was 0.77. Similarly, Liu et al.[33] observed an AUC-ROC of 0.72, and Miotto et al.[36] observed 0.87. Our best overall classification AUC-ROC values from the BOS and BOS-AVR models were 0.88 and 0.92, respectively.

Shah et al.[37] found that AUC-ROC values ranged from 0.70 to 0.76 for predicting a combined outcome of death and cardiovascular hospitalization. Our best overall prediction AUC-ROC values were 0.82, 0.87, and 0.83, respectively from the AVR, BOS, and the BOS-AVR data representations.

Table 2. Number of Unique Representation Records through Representation Mining and Dimensionality Reduction Steps

	Task	Start	Sparsity Screen	Mutual Information	Variable Importance (MDG)
AVR	classification	45,767	6,469	3,000	200
	prediction	25,478	2,552	2,552	
BOS	classification	137,735,403	1,349,704	3,000	200
	prediction	30,962,075	107,760		
BOS-AVR				6,000	200

Please cite this article in press as: Estiri et al., Transitive Sequencing Medical Records for Mining Predictive and Interpretable Temporal Representations, Patterns (2020), https://doi.org/10.1016/j.patter.2020.100051

Table 3. Test Set Median and Best Classification and Prediction Performances across Approaches and Classifiers

		Classification					Prediction				
Classifier	AUC-ROC	AVR	BOS	Δ (%)	BOS-AVR	Δ (%)	AVR	BOS	Δ (%)	BOS-AVR	Δ (%)
Bayesian GLM[a]	median	0.83	0.87	4.4%	0.88	4.4%	0.73	0.79	8.4%	0.67	−8.4%
	best	0.84	**0.88**	4.3%	0.89	4.9%	0.73	0.79	8.4%	0.67	−8.4%
L1 Logistic Reg.[b]	median	0.83	0.85	3%	0.88	5.89%	0.67	0.79	16.7%	0.66	−2.4%
	best	0.86	**0.88**	1.8%	0.89	3.3%	0.72	0.79	9.5%	0.67	−7%
MA Neural Net.[c]	median	0.80	0.85	6.9%	0.85	6.9%	0.67	0.81	20.4%	0.75	12.4%
	best	**0.87**	**0.88**	0.6%	**0.92**	5.4%	0.77	0.85	10.8%	0.79	2.4%
SVM CW[d]	median	0.79	0.80	2.3%	0.83	5.6%	0.79	0.84	6.7%	0.78	−1.4%
	best	0.83	0.83	0%	0.85	2.5%	**0.82**	**0.87**	6.2%	**0.83**	1.3%
Average	median	0.81	0.84	4.1%	0.86	6.2%	0.72	0.81	13%	0.72	0.1%
	best	0.85	0.86	1.5%	0.89	4%	0.76	0.83	8.7%	0.74	−2.9%

Median and best AUC-ROC values are obtained from ten classification iterations with bootstrap cross-validation.
Best AUC-ROC performances are highlighted by bold font.
[a]Bayesian generalized linear model.
[b]Regularized logistic regression (L1).
[c]Model averaged neural network.
[d]Support vector machines with class weights.

The temporal relationships encoded in the BOS approach capture some of the complexities of the clinical process that are lost in the conventional approach. Certain sequences in the BOS model undoubtedly correspond to the clinical narrative that is common for CHF patients. For example, the sequence "abnormalities in breathing → captopril" in the BOS model is a better indicator than either feature on its own in the AVR model (65% versus 49% and 57%). This improved accuracy could be attributable to the sequence's ability to capture the relative timing of the two events. Patients who have abnormalities in breathing may have their symptoms attributable to not just heart failure but also pneumonia, chronic obstructive pulmonary disease, or other pulmonary diseases. Similarly, many patients taking captopril take it for hypertension, chronic kidney disease, or after a myocardial infarction. However, if a patient has abnormalities of breathing and then goes on to take captopril, there is a greater chance that the underlying cause of his or her symptoms and the need for this particular medication is due to CHF. The disease offers a unified explanation for both the symptoms (abnormalities in breathing) and the treatment (captopril). The transitive sequence is better able to represent the patient's clinical experience than either of the individual components on their own.

The sequences that initially seem less obvious could give insight into a disease. For example, as mentioned above, the sequence "topical anti-infectives → unspecified kidney failure" was labeled as difficult to explain. However, an argument could be made that this sequence is still clinically interpretable for heart failure patients. For example, it is likely that patients with heart failure are chronically ill and therefore more susceptible to dermatological infections. Their heart failure could lead to and exacerbate kidney disease. While few physicians would cite this sequence as obvious among this population, it could still be a less recognized but common relationship among heart failure patients. Also, in certain cases such sequences could potentially generate hypotheses for novel clinical relationships not previously appreciated.

Many of the prediction sequences are risk factors for CHF, the corresponding symptoms and medications for those risk factors, or the symptoms of the disease itself. For example, the sequence "essential hypertension → type 2 diabetes mellitus" was identified as a significant sequence. Both features are common and specific risk factors for developing heart failure. It makes sense that such risk factors would be important clinical sequences because they have a known pathological process that leads to the development of heart failure. Moreover, like the classifying sequences, there are also prediction sequences that are less obvious. For example, "gout → encounter for immunization" was identified as an important sequence. At first glance, neither component of this feature seems related to heart failure. However, encounters for immunization are often performed in patients with poorly controlled diabetes mellitus, alcohol use disorder, or cardiovascular disease; all risk factors for heart failure, while risk factors for gout include chronic kidney disease, diabetes mellitus, and cardiovascular disease, all of which are also known risk factors for heart failure. Despite neither feature itself having an obvious direct relation to heart failure, on further analysis both seem likely to have a positive correlation with the disease. The BOS method has the potential to identify predictive sequences that the physician may not otherwise appreciate.

Potential Clinical Utilities of Transitive Sequences

One can envision several potential clinical uses for the transitive sequencing approach. The BOS prediction model can be applied to compute real-time CHF probabilities for all patients who do not have a diagnosis of heart failure. A healthcare provider who is caring for these patients could be given a probability of the patient having the disease based on the available sequences in the chart. This could be of particular value for practicing medicine in under-resourced settings where patients may see healthcare providers less frequently. This tool could help identify patients in a community at risk of developing a particular disease and recommend their evaluation by a healthcare provider. Such predictive algorithms could also assist the

Please cite this article in press as: Estiri et al., Transitive Sequencing Medical Records for Mining Predictive and Interpretable Temporal Representations, Patterns (2020), https://doi.org/10.1016/j.patter.2020.100051

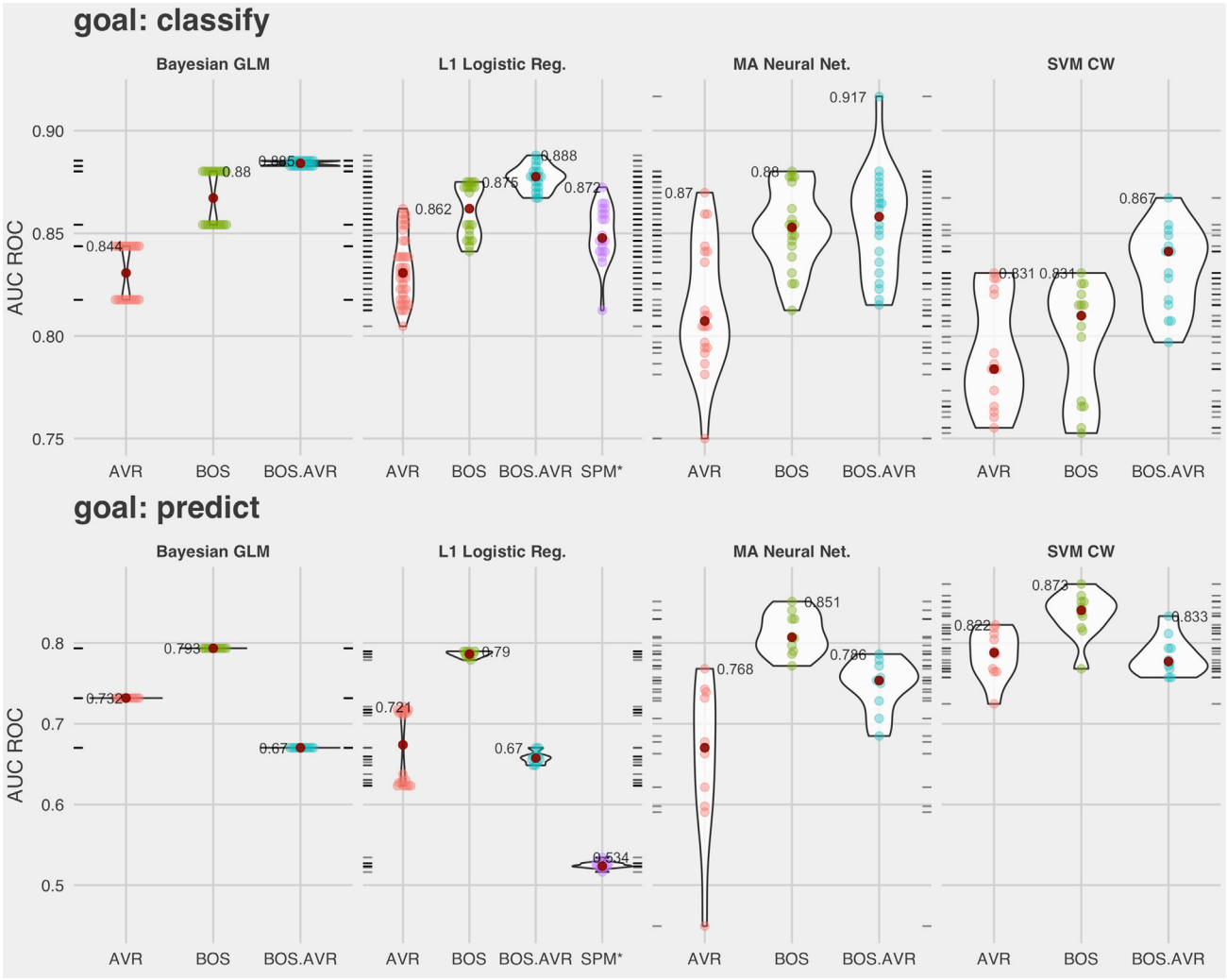

Figure 2. Comparing the AUC-ROC Metrics for Classification and Prediction Tasks by Data Representation

Bayesian GLM, Bayesian generalized linear model; L1 Logistic Reg., regularized logistic regression (L1); MA Neural Net., model averaged neural network; SVM CW, support vector machines with class weights. Median AUC values are identified by red dots, and top values are printed.

provider to consider other diagnoses. There is the potential to use this tool for a variety of different diseases.

Although this is a generic potential use case for any ML algorithms at the point of care, we showed that classification and prediction using transitive sequences have higher accuracy in computing CHF probabilities than what can be computed from raw features.

Sequences from the BOS model can be used as a medication recommender system for patients who have a history of heart failure. The model can provide real-time probabilities for different sequences of diagnoses/medications based on trajectories learned from a large cohort of heart failure patients. For example, our model identified "heart failure → metoprolol" as an important sequence marker for patients with heart failure. Beta-blockers, such as metoprolol, are a standard treatment and have been proved to reduce mortality in CHF patients with a reduced ejection fraction. The BOS model could use these particular sequences as a clinical decision support tool to suggest to providers that they should consider prescribing the medication to their heart failure patients.

Another example is the sequence "atrial fibrillation and flutter → coumarins and indandiones." If a patient with atrial fibrillation has an elevated CHA_2DS_2-VASc score (a standardized score for assessing stroke risk) above a certain threshold, he or she should take an anti-coagulation agent. The score depends on various pre-existing conditions, one of which is CHF. The sequencing approach would recommend anti-coagulation after atrial fibrillation if the patient has a CHF record in his or her medical history. Such a clinical decision support tool could be especially useful for generating recommendations for patients with complex histories, multiple providers, and health records that span many years.

Finally, the use of classification sequences in cohort identification can have utility in large patient cohorts. They can more accurately identify appropriate patients for clinical trials, quality assessment, and biomedical research. For example, if a

Please cite this article in press as: Estiri et al., Transitive Sequencing Medical Records for Mining Predictive and Interpretable Temporal Representations, Patterns (2020), https://doi.org/10.1016/j.patter.2020.100051

CellPress
OPEN ACCESS

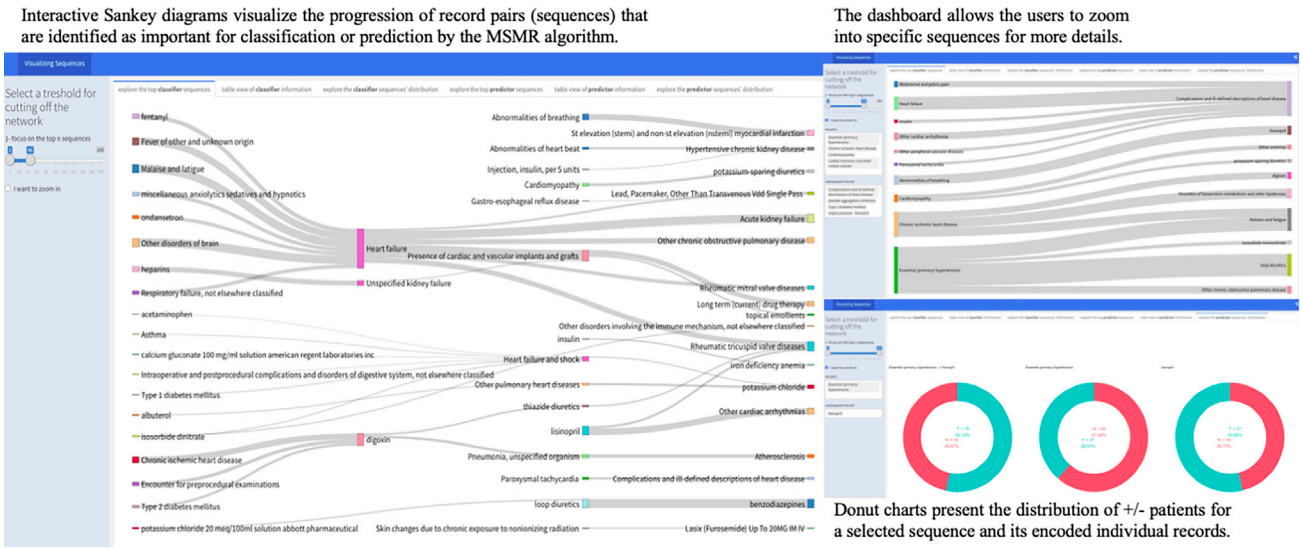

Figure 3. A Snapshot of the Developed Graphical Dashboard for Further Exploration of the Sequences

researcher wanted to rapidly select all patients with heart failure in a given population, this could be done with greater accuracy using transitive sequences and the BOS model.

Relationship to Recurrent Neural Networks

To some extent, this study can be a simplified reverse engineering of recurrent neural networks (RNNs).[38] Instead of searching for n-deep sequences, we use the shortest sequences and perform dimensionality reduction. RNNs and RNN-based models such as long short-term memory (LSTM)[39] and gated recurrent unit (GRU)[40] have been recently applied to clinical questions to account for time.[41–44] However, the challenge of applying RNN-based algorithms to EHR data is twofold. On the one hand, there is a tradeoff between accuracy and interpretability that needs to be carefully considered. More complex algorithms often result in highly accurate models but are difficult to understand.[45] Although ways for making sense of RNN-based models are being explored, real application in clinical care, whereby both accuracy and interpretability are critical, is still a long way away. On the other hand, discrete healthcare records in EHRs often do not precisely reflect the true health status of a patient. Expecting a set of recurrent layers to somehow make sense of the data points that may or may not be reliable is naive.

Limitations and Future Work

A limitation of this work is that we did not filter disease observations by their phenotypic expression patterns. Patients' health states evolve over diverse time scales. Acute diseases such as pneumonia are more isolated spontaneous occurrences, while chronic conditions such as diabetes develop and progress over years. Acute conditions tend to have lower

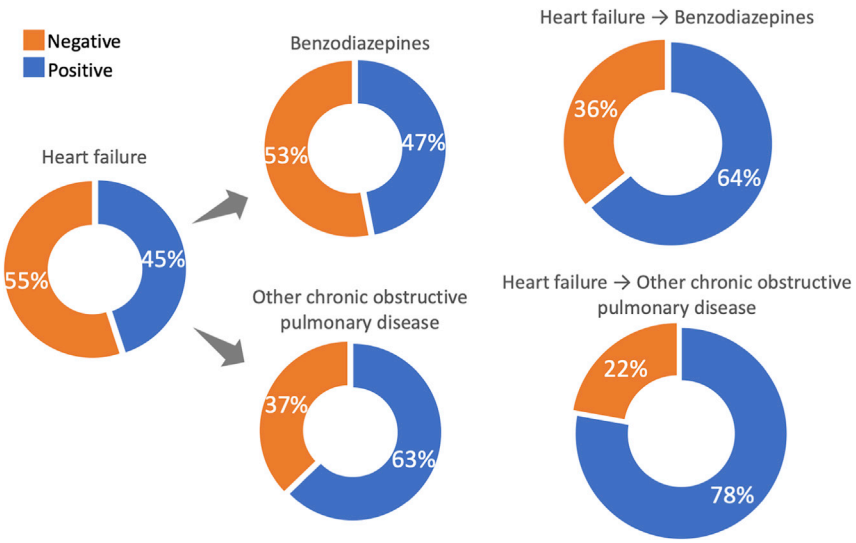

Figure 4. Phenotype Probability by Data Representation

The donut charts illustrate the probability of phenotype for patients who have at least one record of the feature. AVR data representations are common medication or diagnosis codes in EHRs. The BOS data representations are sequenced features mined in this research.

Table 4. Illustrative Examples of Expected and Obscure Transitive Sequences

Easy to Explain	Difficult to Explain
Classification	
• abnormalities of breathing → cardiomyopathy • abnormalities of breathing → captopril • magnesium gluconate → cardiomyopathy • heart failure → pleural effusion • heart failure → metoprolol	• topical anti-infectives → unspecified kidney failure • dorsalgia → cardiomyopathy • GERD → pacemaker • metoprolol → levofloxacin • nail disorders → furosemide
Prediction	
• essential hypertension → lisinopril • chronic ischemic heart disease → angina pectoris • essential hypertension → type 2 diabetes mellitus • cardiomyopathy → platelet aggregation inhibitors • cardiac murmur → complications of heart disease	• docusate → SSRI anti-depressants • docusate → propofol • ondansetron → glucose • fever of unknown origins → vancomycin • vancomycin → fentanyl

GERD, gastroesophageal reflux disease; SSRI, selective serotonin reuptake inhibitor.

entropies, indicating an inherent link between the predictability of disease and their phenotypic expression pattern.[46] Scattered in EHRs are records of acute conditions, which often do not exhibit long-range patterns. Therefore, filtering acute conditions out may improve the temporally correlated predictive power.

Also, visual dashboard development was not a primary objective of this study. While our preliminary findings suggest that a visual dashboard can be useful in explaining the complex relationships to clinicians, a formal user study and further evaluations are needed to develop an interactive visual interface.

Finally, our modeling primarily focused on CHF. Further research is needed to evaluate the performance of transitive sequences for classifying/predicting other diseases and to further explore the possibilities of incorporating the sequencing approach into decision support tools at the point of care.

Conclusion

Innovative methods that enable us to properly incorporate time and understand the complexities involved in the healthcare process can yield interpretable findings from large-scale clinical databases. We found that data representations mined from sequences of EHR events are better phenotype "differentiators" and predictors than the "atemporal" EHRs that are widely used as the primary data representations in ML.

Given the rapidly increasing prevalence of EHR systems in today's practice, exploiting the temporal information in EHRs can advance medical knowledge discovery and meaningfully change clinical care by identifying and validating novel disease markers. The transitive sequencing approach presented here allows for limited expert involvement in feature engineering. However, the graphical experiment was helpful in that it resulted in refining the important sequences. Much like the genomics community, the identified sequences of medical records can be cataloged and shared on an accessible platform that would allow for the collaborative clinical use of the sequences as risk factors for diseases in many domains.

EXPERIMENTAL PROCEDURES

Resource Availability

Lead Contact

Hossein Estiri, PhD. hestiri@mgh.harvard.edu.

Materials Availability

This study did not generate any new unique reagents or materials.

Data and Code Availability

Protected Health Information restrictions apply to the availability of the clinical data here, which were used under Institutional Review Board approval for use only in the current study. As a result, these datasets are not publicly available. All code used for modeling and dimensionality reduction in this study uses open-source R packages. The transitive sequencing code is available on https://github.com/hestiri/TSPM under Mozilla Public License 2.0.

Study Design

Modeling the temporal information in clinical data can uncover other dimensions of healthcare delivery that generate signals for disease classification or prediction.[3] Thus, we hypothesize that temporal sequences of electronic medical records will improve (1) computational classification of patient cohorts and (2) phenotype prediction. To test these hypotheses, we construct a set of baseline models applying the conventional approach of aggregating the frequency of medical events as features for downstream ML algorithms. We also mine a set of representations by transitive sequencing of the medication and diagnosis events in electronic medical records. We call this proposed approach the BOS approach. The study design can be characterized under representation mining, dimensionality reduction, and ML experiments, and is illustrated in Figure 1.

Data

To test the study hypotheses, we used EHR data from the Mass General Brigham Biobank. Data from all Biobank patients with at least an International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) code for CHF (428.0) were included in this study. This cohort consisted of 6,857 patients consented into the Partners Biobank. To create a gold-standard dataset, a board-certified nurse reviewed the clinical notes of a random sample of these patients. The review resulted in 56 patients with gold-standard CHF labels. Specifically, for each of these patients in the gold-standard dataset we had curated labels that determined whether the patient truly had (or did not have) heart failure. We used data from the 56 patients for testing. For training, we used the data from the remaining cohort of 6,851 Biobank patients. Table 1 provides summary information about the testing and training data. The training data included approximate training labels (i.e., silver-standard labels) curated by validated algorithms,[47,48] which use the distributional properties of a small number of representative features in the gold-standard population to estimate disease probabilities for all patients. The silver-standard labels were not verified by human experts but are crucial for scaling up ML training on large-scale clinical data. The use of data for this study was approved by the Mass General Brigham Institutional Review Board (2017P000282).

We only used the medications and diagnoses records data. For the diagnosis records, we used the ICD-9/10-CM. For medications, we used RxNorm codes.

For classification, we utilized the entire data available for patients in the cohort; therefore, data from potentially before and after CHF diagnoses codes are used. The prediction data is a subset of the classification data, in which we only extracted medication and diagnosis records from before the first time a record of heart failure was observed in a patient's medical records. For a number of patients, the medical records began with a CHF diagnosis. Hence, the cohort size in the prediction dataset is slightly smaller. As a result of the way we defined the datasets, none of the patients remaining in the data for training prediction classifiers

had a record of CHF elsewhere in their medical records. There is no specific time frame for prediction. In the prediction dataset, the next encounter recorded in the electronic medical records for a patient will include a CHF diagnosis.

Representation Mining

To test the study hypotheses that transitive sequences of electronic medical records can improve computational classification of patient cohorts and phenotype prediction, we mined two vectors of data representations. The study design is illustrated in Figure 1. We first constructed a baseline method that applies the conventional approach for using EHR observations as features for phenotype classification and prediction. We henceforth call this the AVR approach. The frequency of all medical events is counted for each patient in the EHR data, and the patient is represented by a vector of the length equal to the number of unique events in her or his medical records.

Given a list $O_1, O_2, \ldots, O_n$ of diagnosis or medication observations, for each patient p, we have recorded the times $t^p_{i1} \leq t^p_{i2} \leq \ldots \leq t^p_{ik^p_i}$ at which the observation O_i was recorded (we allow $k^p_i = 0$, in which case observation O_i was yet to be recorded for patient p).

AVR Representations

In the AVR approach, our features are the possible observations, and for each patient, say patient p, we record only the numbers $k^p_1, k^p_2, \ldots, k^p_n$ of records of each observation. For each i, we think of the k^p_1's as samples of a random variable X_i. Our goal is then to predict the class label Y, given $X_1, X_2, \ldots, X_n$.

We also mined a set of representations by transitive sequencing of the medication and diagnosis events in electronic medical records. In this approach, the patient is represented by a vector of the length equal to the number of sequences in her/his medical records. We call this proposed approach the BOS approach.

BOS Representations

In the BOS approach, the features are all possible pairs of distinct observations (O_i, O_j), $i \neq j$. For a fixed patient p, and $i \neq j \leq n$, we set r_{ijp} to be 1 if $k^p_i \geq 1$, $k^p_j \geq 1$, and $t^p_{i1} \leq t^p_{j1}$, and 0 otherwise. In words, r_{ijp} is 1 if and only if both O_i and O_j were recorded for the patient, and the first record of observation i was before, or at the same time as, the first record of observation j. Again, for each fixed $i \neq j$, we think of the r_{ijp}s as samples of a random variable X_{ij}. Our goal is then to predict the class label Y given. $(X_{ij})_{i \neq j}$.

The use of first record (rather than all records) is a major difference in the way we mined sequences when compared with SPM routines in the general data mining community. We felt this choice better reflects the actual patient state to handle the repeated problem list entries for two reasons. First, diagnosis records are generally carried forward in patients' medical records (often known as problem list entries) through all succeeding encounters, making the first occurrence the true start of a sequential pattern. Second, relying on the high-resolution timing data of repeated diagnosis records is an implementation detail of the clinical system rather than clinically meaningful evidence of the patient's medical history.

It is important to emphasize that we call the sequential pairs in the BOS approach transitive sequences, as they embody distinctive modifications to the conventional SPM. Imagine a sequential pattern where observation A happened directly before B, and B happened directly before C ($A \rightarrow B \rightarrow C$). SPM mines subsequences $A \rightarrow B$ and $B \rightarrow C$. To account for the potential biases in EHRs, the transitive sequencing algorithm mines subsequences $A' \rightarrow B'$, $B' \rightarrow C'$, but also $A' \rightarrow C'$ from the sequence $A' \rightarrow B' \rightarrow C'$, where A', B', and C' are the first records of A, B, C in the medical records. To evaluate the performance difference between the BOS approach and the SPM, we also mined the SPM sequences and used the most frequent sequences for classification.

Dimensionality Reduction

We apply a form of entropy-based temporal representation mining of discrete events from clinical data, which deviates from the traditional SPM and TA approaches that use frequency-based criteria for selecting subsequences. If all pairs of sequences in the BOS approach exist, there will be exactly $\frac{n(n-1)}{2}$ pairs (i,j) with $i \neq j$ and $i,j \leq n$. Thus, the number of sequential features is roughly quadratic in the number of observations. As demonstrated in Results, the sequence mining resulted in the explosion of sequences and therefore left us with a highly dimensional vector of representations. To both the BOS and AVR representations, we applied the MSMR formal dimensionality reduction procedure.

To minimize sparsity, we removed any feature that has prevalence smaller than 1%. On the remaining features, we compute the empirical mutual information using an estimation of the entropy of the empirical probability distribution.[49,50] Mutual information provides a measurement of the mutual dependence between two random variables, which unlike most correlation measures can capture non-linear relationships.[50,51] We ranked the data representations based on their mutual information with the labeled outcome (in ties, we used prevalence to determine the ranking) and conventionally selected the top 3,000 representations from the AVR and BOS approaches.

We further scrutinized the relevance through random forests (RF)[52] using the MDG—also known as Gini importance—for variable importance. The Gini importance measures the node purity gain by splitting a variable.[53] A variable's MDG is a forest-wide weighted average of the decrease in the Gini Impurity metric resulting from splitting on the variable across all of the individual trees that make up the forest.[54] A higher MDG indicates higher variable importance. At the end of this step, using the median MDGs we ranked features and conveniently curated feature sets for each approach containing the top 200 features. We also combined the two feature vectors prior to the MDG computation step and computed MDGs for the combined data representations as a hybrid approach (AVR-BOS). At the conclusion of the MSMR procedure, we had curated three feature sets through the MSMR procedure containing the top 200 AVR, BOS, and BOS-AVR representations. We also added the top-200 frequent sequences to represent the conventional frequent SPM approach.

Training Classification and Prediction Classifiers

We trained four different classifier algorithms on each vector of data representations using bootstrap cross-validation: (1) logistic regression with L1 regularization; (2) Bayesian generalized linear model; (3) model averaged neural network; and (4) support vector machines with class weights.[55] For SPM we only trained the regularized logistic regression (L1) classifier. All variables were scaled and centered. All classifiers were trained and tested on the 6,851-patient data. All performance metrics were computed using the 56-patient test set. Furthermore, we iterated the training process ten times with bootstrap sampling and used the median performance metrics for comparing the approaches. Overall, for each of the approaches, we trained 40 classifiers (4 algorithms × 10 bootstrap sampling iterations).

Evaluation

To evaluate our two hypotheses, we compared classifier performances using the AUC-ROC curve. We applied non-parametric and post hoc tests for comparing and ranking the classification performances across the 120 classifiers (40 classifiers × 3 approaches). The goal was to evaluate whether the AUC-ROC values would provide enough statistical evidence that the classifiers have different performances.

Finally, we developed an interactive graphical dashboard to evaluate the important sequential patterns discovered through the dimensionality reduction process. The dashboard provided different visualization and table views of the top 200 transitive sequences using the RStudio Shiny platform. The dashboard design involved an iterative process with a small number of physicians in our group to adjust the visualizations/tables and add new one for facilitating their interpretation of the sequences. Using the graphical dashboard, the top 200 BOS transitive sequences were evaluated by the physicians for their clinical significance. Specific sequences that were identified as having clinical meaning were evaluated further based on their accuracy for identifying patients with heart failure.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.patter.2020.100051.

ACKNOWLEDGMENTS

This work was funded through the U.S. National Human Genome Research Institute grant R01-HG009174.

AUTHOR CONTRIBUTIONS

Conceptualization, H.E., T.H.M., and S.N.M.; Methodology, H.E.; Formal Analysis, H.E.; Investigation, H.E., S.N.M., and Z.H.S.; Writing – Original Draft, H.E. and Z.H.S.; Writing – Review & Editing, H.E., Z.H.S., J.G.K., K.B.W., and S.V.; Visualization, H.E.; Data Curation, V.M.C. and M.E.M.; Funding Acquisition, S.N.M.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 13, 2020
Revised: April 27, 2020
Accepted: May 26, 2020
Published: June 18, 2020

REFERENCES

1. Hripcsak, G., Albers, D.J., and Perotte, A. (2011). Exploiting time in electronic health record correlations. J. Am. Med. Inform. Assoc. *18* (*Suppl 1*), i109–i115.
2. Hripcsak, G., and Albers, D.J. (2013). Next-generation phenotyping of electronic health records. J. Am. Med. Inform. Assoc. *20*, 117–121.
3. Agniel, D., Kohane, I.S., and Weber, G.M. (2018). Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ *361*, https://doi.org/10.1136/bmj.k1479.
4. Moskovitch, R., Choi, H., Hripcsak, G., and Tatonetti, N. (2017). Prognosis of clinical outcomes with temporal patterns and experiences with one class feature selection. IEEE/ACM Trans. Comput. Biol. Bioinform. *14*, 555–563.
5. Dahlem, D., Maniloff, D., and Ratti, C. (2015). Predictability bounds of electronic health records. Sci. Rep. *5*, 11865.
6. Madkour, M., Benhaddou, D., and Tao, C. (2016). Temporal data representation, normalization, extraction, and reasoning: a review from clinical domain. Comput. Methods Programs Biomed. *128*, 52–68.
7. Batal, I., Valizadegan, H., Cooper, G.F., and Hauskrecht, M. (2013). A temporal pattern mining approach for classifying electronic health record data. ACM Trans. Intell. Syst. Technol. *4*, https://doi.org/10.1145/2508037.2508044.
8. Liu, Z., Wu, L., and Hauskrecht, M. (2013). Modeling clinical time series using Gaussian process sequences. In Proceedings of the 2013 SIAM International Conference on Data Mining.
9. Albers, D.J., and Hripcsak, G. (2012). Estimation of time-delayed mutual information and bias for irregularly and sparsely sampled time-series. Chaos Solitons Fractals *45*, 853–860.
10. Zhao, J., Papapetrou, P., Asker, L., and Boström, H. (2017). Learning from heterogeneous temporal data in electronic health records. J. Biomed. Inform. *65*, 105–119.
11. Pivovarov, R., and Elhadad, N. (2015). Automated methods for the summarization of electronic health records. J. Am. Med. Inform. Assoc. *22*, 938–947.
12. Stacey, M., and McGregor, C. (2007). Temporal abstraction in intelligent clinical data analysis: a survey. Artif. Intell. Med. *39*, 1–24.
13. Orphanou, K., Stassopoulou, A., and Keravnou, E. (2014). Temporal abstraction and temporal Bayesian networks in clinical domains: a survey. Artif. Intell. Med. *60*, 133–149.
14. Shahar, Y., and Musen, M.A. (1993). RÉSUMÉ: a temporal-abstraction system for patient monitoring. Comput. Biomed. Res. *26*, 255–273.
15. Shahar, Y., and Musen, M.A. (1996). Knowledge-based temporal abstraction in clinical domains. Artif. Intell. Med. *8*, 267–298.
16. Moskovitch, R., and Shahar, Y. (2015). Classification-driven temporal discretization of multivariate time series. Data Min. Knowl. Discov. *29*, 871–913.
17. Moskovitch, R., Walsh, C., Hripsack, G., and Tatonetti, N. (2014). Prediction of biomedical events via time intervals mining. J. Biomed. Inform. *75*, 70–82.
18. Batal, I., Cooper, G., Fradkin, D., Harrison, J., Moerchen, F., and Hauskrecht, M. (2016). An efficient pattern mining approach for event detection in multivariate temporal data. Knowl. Inf. Syst. *46*, 115–150.
19. Shknevsky, A., Shahar, Y., and Moskovitch, R. (2017). Consistent discovery of frequent interval-based temporal patterns in chronic patients' data. J. Biomed. Inform. *75*, 83–95.
20. Bellazzi, R., Larizza, C., and Riva, A. (1998). Temporal abstractions for interpreting diabetic patients monitoring data. Intell. Data Anal. *2*, 97–122.
21. Bellazzi, R., Larizza, C., Magni, P., Montani, S., and Stefanelli, M. (2000). Intelligent analysis of clinical time series: an application in the diabetes mellitus domain. Artif. Intell. Med. *20*, 37–57.
22. Agrawal, R., Srikant, R., and Others. (1995). Mining sequential patterns. icde *95*, 3–14.
23. Fournier-Viger, P., Lin, J.C.-W., Kiran, R.U., Koh, Y.S., and Thomas, R. (2017). A survey of sequential pattern mining. Data Sci. Pattern Recognit. *1*, 54–77.
24. Truong-Chi, T., and Fournier-Viger, P. (2019). A survey of high utility sequential pattern mining. In High-Utility Pattern Mining, P. Fournier-Viger, J.C.-W. Lin, R. Nkambou, B. Vo, and V. Tseng, eds., pp. 97–129.
25. Mabroukeh, N.R., and Ezeife, C.I. (2010). A taxonomy of sequential pattern mining algorithms. ACM Comput. Surv. *43*, https://doi.org/10.1145/1824795.1824798.
26. Berlingerio, M., Bonchi, F., Giannotti, F., and Turini, F. (2007). Mining clinical data with a temporal dimension: a case study 2007 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2007), 429–436.
27. Moskovitch, R., and Shahar, Y. (2015). Classification of multivariate time series via temporal abstraction and time intervals mining. Knowl. Inf. Syst. *45*, 35–74.
28. Batal, I., Sacchi, L., Bellazzi, R., and Hauskrecht, M. (2009). A temporal abstraction framework for classifying clinical temporal data. AMIA Annu. Symp. Proc. *2009*, 29–33.
29. Moskovitch, R., Polubriaginof, F., Weiss, A., Ryan, P., and Tatonetti, N. (2017). Procedure prediction from symbolic Electronic Health Records via time intervals analytics. J. Biomed. Inform. *75*, 70–82.
30. Garofalakis, M.N., Rastogi, R., and Shim, K. (1999). SPIRIT: sequential pattern mining with regular expression constraints. VLDB *99*, 7–10.
31. Zaki, M.J. (2001). Parallel sequence mining on shared-memory machines. J. Parallel Distrib. Comput. *61*, 401–426.
32. Ayres, J., Flannick, J., Gehrke, J., and Yiu, T. (2002). Sequential PAttern mining using a bitmap representation. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 429–435.
33. Liu, C., Wang, F., Hu, J., and Xiong, H. (2015). Temporal phenotyping from longitudinal electronic health records. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15). 10.1145/2783258.2783352.
34. Guo, S., Li, X., Liu, H., Zhang, P., Du, X., Xie, G., and Wang, F. (2017). Integrating temporal pattern mining in ischemic stroke prediction and treatment pathway discovery for atrial fibrillation. AMIA Jt. Summits Transl. Sci. Proc. *2017*, 122–130.
35. Wu, J., Roy, J., and Stewart, W.F. (2010). Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med. Care *48*, S106–S113.
36. Miotto, R., Li, L., Kidd, B.A., and Dudley, J.T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci. Rep. *6*, https://doi.org/10.1038/srep26094.
37. Shah, S.J., Katz, D.H., Selvaraj, S., Burke, M.A., Yancy, C.W., Gheorghiade, M., Bonow, R.O., Huang, C.-C., and Deo, R.C. (2015). Phenomapping for novel classification of heart failure with preserved ejection fraction. Circulation *131*, 269–279.

38. Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning representations by back-propagating errors. Nature *323*, 533–536.

39. Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural Comput. *9*, 1735–1780.

40. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv, 1406.1078.

41. Wu, S., Liu, S., Sohn, S., Moon, S., Wi, C.-I., Juhn, Y., and Liu, H. (2018). Modeling asynchronous event sequences with RNNs. J. Biomed. Inform. *83*, 167–177.

42. Lipton, Z.C., Kale, D.C., Elkan, C., and Wetzel, R.C. (2015). Learning to diagnose with LSTM recurrent neural networks. arXiv, 1511.03677.

43. Pham, T., Tran, T., Phung, D., and Venkatesh, S. (2017). Predicting healthcare trajectories from medical records: a deep learning approach. J. Biomed. Inform. *69*, 218–229.

44. Choi, E., Bahadori, M.T., Song, L., Stewart, W.F., and Sun, J. (2017). GRAM: graph-based attention model for healthcare representation learning. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 787–795.

45. Choi, E., Bahadori, M.T., Sun, J., Kulas, J., Schuetz, A., and Stewart, W. (2016). RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. In Advances in Neural Information Processing Systems, *Vol. 29*, D.D. Lee, M. Sugiyama, U. V Luxburg, I. Guyon, and R. Garnett, eds. *(Curran Associates, Inc.), pp. 3504–3512.*

46. Perotte, A., and Hripcsak, G. (2013). Temporal properties of diagnosis code time series in aggregate. IEEE J. Biomed. Heal. Inform. *17*, 477–483.

47. Yu, S., Ma, Y., Gronsbell, J., Cai, T., Ananthakrishnan, A.N., Gainer, V.S., Churchill, S.E., Szolovits, P., Murphy, S.N., Kohane, I.S., et al. (2018). Enabling phenotypic big data with PheNorm. J. Am. Med. Inform. Assoc. *25*, 54–60.

48. Wagholikar, K.B., Estiri, H., Murphy, M., and Murphy, S.N. (2020). Polar labeling: silver standard algorithm for training disease classifiers. Bioinformatics *36*, 3200–3206.

49. Meyer, P.E. (2008). Information-Theoretic Variable Selection and Network Inference from Microarray Data, Ph. D. Thesis (Univ. Libr. Bruxelles).

50. Cover, T.M., and Thomas, J.A. (2012). Elements of Information Theory (John Wiley & Sons).

51. Paninski, L. (2003). Estimation of entropy and mutual information. Neural Comput. *15*, 1191–1253.

52. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

53. Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P. (2013). Understanding variable importances in forests of randomized trees. 26th International Conference on Advances in Neural Information Processing Systems (NIPS 2013) 431–439

54. Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., and Hamprecht, F.A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics *10*, 213.

55. Batuwita, R., and Palade, V. (2013). Class imbalance learning methods for support vector machines. In Imbalanced Learning: Foundations, Algorithms, and Applications, H. He and Y. Ma, eds. (IEEE/Wiley), pp. 83–99.