High-throughput Phenotyping with Temporal Sequences

Hossein Estiri[#], Thomas H. McCoy, Jr, Kavishwar B. Wagholikar, Alyssa P. Goodson, Katie Murphy, Shawn N. Murphy

Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA, USA.
Center for Quantitative Health, Massachusetts General Hospital, Boston, MA, USA.
Research Information Science and Computing, Partners HealthCare, Charlestown, MA, USA.
Harvard Medical School, Boston, MA, USA.

[#]Corresponding Author: Hossein Estiri hestiri@mgh.harvard.edu

ABSTRACT

Electronic health records (EHRs) contain important temporal information about progression of disease and treatment outcomes. The objective of this paper is to propose and test a high-throughput approach for phenotyping using temporal sequences of EHR observations to derive predictive and interpretable data representations. Using expert labeled clinical data from a congestive heart failure (CHF) cohort, we applied a four-phase framework to mine new data representations from temporal sequences of EHR observations, reduce dimensionality of the feature space, evaluate feature importance, and classify CHF patients with CHF. We found that sequenced features significantly outperform raw EHR features in classifying heart failure patients. Compared to the gold standard created by a panel of domain experts, the best classifier using sequences of diagnosis and medication observations classified heart failure status with a receiver operating characteristic area under the curve of 0.926. Compared with their raw counterparts, the top sequenced features have lower prevalence in medical records, but present more useful information for classification as the majority of those who have the sequences are likely to have the disease. Raw EHR data are not direct reflections of true health states as they also reflect various healthcare processes. Our results demonstrated that data representations obtained from sequencing EHR observations can present novel insight about the progression of disease that is difficult to discern when clinical observation histories are treated independently. Harnessing knowledge of disease progression through temporal sequencing of EHR observations can improve computational disease classification or phenotyping.

INTRODUCTION

The widespread adoption of electronic health records (EHRs) has led to an unprecedented load of patient health information in electronic format. Despite the natural excitement emerging from the large amount of information presented by EHRs, daunting challenges remain. The overwhelming availability of electronic medical records has come at the cost of increasing physician burnout and has raised concerns about information overload. Dimensionality, sparsity, heterogeneity, and quality issues present significant impediments for secondary use of EHR data.[1,2] In addition, as the primary impetus for EHR implementation has been clinical care, EHR observations reflect a complex set of processes that further obscure their utility in research. The raw EHR data are not direct indicators of a patient's true health state, but rather reflect the clinical processes (e.g., policies and workflows of the provider and payor organizations), patient's interaction with the system, and the recording process.[3–5] Hripcsak et al. (2011) describe this as the healthcare process model.[3]

If we treat EHR data as reflective only of the physiology of the patients, we will miss out other attributes that are embedded in EHR observations and can uncover many more dimensions of the healthcare delivery (e.g., clinician's decision making process) and disease progression, and generate new signals for classification or prediction.[5] The inherent complexities introduce additional bias to EHR data, warranting fundamentally different approaches to analyze EHR observations and interpret their results.[4] In this context, the temporal nature of observational data stored in EHRs is perplexing and has not been fully exploited by current methods. [6] Electronic health records contain important temporal that presented a significant opportunity to apply innovative data mining methods and discover important medical knowledge about disease progression and patients.[7–9] The temporal properties of EHR data signify the complexities involved in uncovering the healthcare processes and the challenges to properly incorporate temporal information in generalizable analytical techniques.[10]

Despite a long tradition in informatics, translation of data mining methodologies to the complexity of the clinical workflow is slow moving. Biomedical researchers increasingly apply conventional association studies to EHR data, yet the temporality of these data has not been fully exploited.[3] The conception of time, in general, can be domain-specific and complex, entailing significant domain knowledge to accurately understand and harness.[11] The time-stamped EHR observations, in particular, are often acquired asynchronously (i.e., measured at different time instants and sampled irregularly in time), sparse, and include heterogeneous longitudinal data (often both time points and time intervals), thus provide fundamental challenges for directly applying common temporal analysis methods.[11–16] For example, the record dates associated with diagnoses codes in EHRs often reflect when the diagnoses were made by clinicians, not the actual onset of the disease.

More complex algorithms such as Recurrent Neural Networks (RNNs)[17] and RNN-based models such as Long Short-Term Memory (LSTM)[18] and Gated Recurrent Unit (GRU)[19] have been used to account for time.[20–27] These algorithms often result in highly predictive models, but they are hard to understand, limiting their utility in healthcare settings. Nevertheless, accurate and interpretable temporal representation of clinical observations in EHR data remains an open problem.[28]

This is a critical research gap, as properly incorporating temporality of EHR observations can present new signals and yield to interpretable findings from large scale clinical databases.[3,5] Innovative methods that enable us to properly incorporate time and understand complexities involved in healthcare process can yield to interpretable findings from large scale clinical databases.[3] The increasing prevalence of EHR

systems provide an outstanding opportunity to design novel approaches to mining features from EHR observations that can represent complexities of healthcare processes while improving classification/prediction and interpretability.

A primary objective of precision medicine is to develop computational methods that can be utilized in predicting health status, disease, or disability for patients.[29] Leveraging EHRs to identify patient cohorts for clinical and genomic research has become increasingly popular. EHR-based phenotyping utilizes the information in a patient's health records to infer presence of a disease (or lack thereof).[30] Despite the fact that developing phenotypic definitions from EHR data is still expensive, requiring significant involvement by domain experts,[29,31–33] many healthcare institutions[34,35,44–48,36–43] are actively involved in constructing and validating EHR phenotyping algorithms. Feature engineering has large impacts on the success of classification (e.g., phenotyping) and prediction algorithms. The goal of feature selection is to improve prediction performance, reduce prediction cost, and offer a better understanding of the underlying patterns that generated the data.[49] Efforts to curate computational phenotypes and discover clinical knowledge from EHR observations must account for the potential biases introduced through the recording process.[4]

In this paper, we focus on feature engineering for curating computational phenotypes. Considering electronic health records as 'indirect' reflections of a patient's true health state, we propose a high-throughput approach for phenotyping using temporal sequences of EHR observations to derive predictive and interpretable data representations. We show that sequenced features significantly outperform raw EHR features in classifying heart failure patients and present novel insight that are difficult to discern from raw EHRs observations.

BACKGROUND AND SIGNIFICANCE

Biomedical research increasingly applies advanced computational techniques to clinical data in order to identify disease comorbidities, stratify patient cohorts, study drug interactions, and predict disease.[1] While interest in this approach is growing, progress is inhibited by complexities of adequately representing time in EHR observations. Only recently have investigators begun developing and applying methods to extract temporal knowledge from EHR data.

Clinical observations in EHRs are often stored at multiple temporal granularities, requiring transformation into a more uniform dimension for data mining.[50] Temporal abstraction can derive symbolic representations of clinical observations captured in EHR data.[9] Temporal abstraction uses domain knowledge to transform multivariate temporal data into a set of symbolic interval-based presentation.[12,50,51] In temporal abstraction, heterogeneous temporal data are processed into homogeneous representations providing means to achieve precise data descriptors that can be used as input to a reasoning engine to describe complex temporal patterns and relationships among clinical variables and concepts.[50,52,53] Interest in temporal abstraction has been growing in biomedical research [54–62] – for comprehensive surveys on temporal abstraction literature in the clinical domains see [52] and [63].

Naturally, much of the prior research applying some temporal abstraction to EHR data has happened after 2009. For example, Hripcsak et al. (2009) used temporal abstraction to measure the uncertainty of temporal assertions in clinical texts.[64] Savova et al. (2009) applied temporal abstraction to study temporal relation and timeline discovery from the clinical narrative.[8] Xu et al. (2013)created an end-to-

end system to identify temporal relations in clinical discharge summaries.[65] Moskovitch et al. (2014 & 2017) used temporal abstraction to transform the data into symbolic time intervals series and applied a fast time intervals mining algorithm to discover frequent Time Intervals Related Patterns (TIRPs) that were used as features to predict frequent and clinically important procedures.[9,15] Jane et al. (2016) Applied a temporal rough set induced neuro-fuzzy (TRiNF) mining framework for temporal data acquisition and temporal disease classification.[66] Cheng et al. (2017) applied temporal abstraction to extract vital features for predicting the progression of Chronic Kidney Disease.[67]

Accounting for the temporality of clinical data through various temporal abstractions has repeatedly exhibited improved predictive modeling performance and efficiency over, for example, atemporal methods in predicting clinical procedures,[6,9] hospital admission orders, [68], Chronic Kidney Disease [67], and loss of kidney function [7], and popular machine learning and deep learning approaches for early diagnosis tasks for septic shock.[69]

Efforts to curate computational phenotypes and discover clinical knowledge from EHR observations must account for the potential biases introduced through the recording process.[4] Feature engineering plays a key role in the performance and interpretability of computational algorithms. Currently, engineering clinically meaningful features relies on a heavy dose of domain expert involvement, using complex ad-hoc procedures[70] that are often neither generalizable nor scalable, limiting opportunities to discover novel patterns from data.[29]

APPROACH

EHR phenotyping is transforming the raw EHR data into clinically-relevant features.[4] The conventional approach to extracting features from EHR observations for prediction or classification is to obtain patient level observation counts over a certain period of time and aggregating them to obtain a cohort level. Features can be all or a selected set of raw clinical observations. Temporal patterns are critical in developing EHR-based phenotypes.[71] Temporal abstraction is being increasingly applied in classification of complex multivariate time series data.[12] Utility of the mined temporal patterns as features for predicting and/or classifying various health outcomes has been demonstrated.[15] Our approach to extracting features is through sequencing EHR observations (Figure 1). We hypothesize that sequences of EHR observations can offer more useful information about healthcare processes and therefore are better features for computational disease classification (phenotyping) and prediction using EHR data.

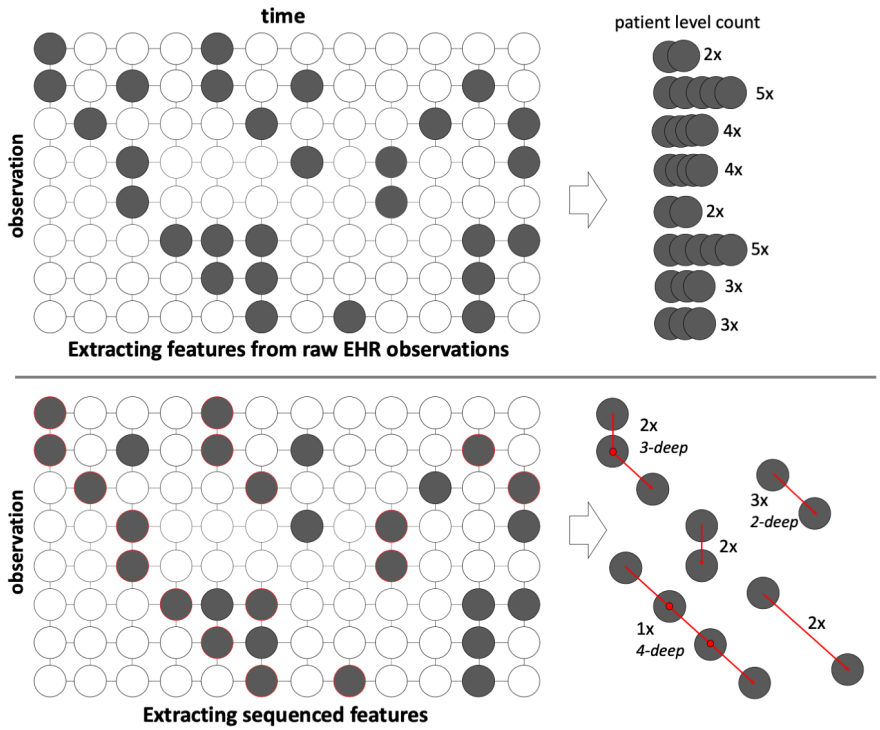


Figure 1. Comparing the conventional and proposed approaches to extracting sequenced features from raw EHR observations.

Following the healthcare process model, in this approach, we consider electronic health records as indirect reflections of a patient's true health state. These sequenced features provide a form of temporal abstraction to study EHR observations and their interactions, under the rubric of a complex network of time-ordered clinical observations. According to network science, such simplifications of complex network interactions in network often can result in useful discoveries.[72]

Experimental Design

Moskovitch and Shahar (2015) presented a three-phase framework for classification of multivariate time series: (1) temporal abstraction to extract a series of symbolic time intervals, (2) mining the time intervals to discover frequent patterns, and (3) using the patterns as features to develop a classifier.[12] Mining temporal features usually return a large number of temporal patterns, most of which are often irrelevant to the classification task.[13,73] In the context of phenotyping with usually small gold-standard training sets, a large number of temporal patterns can aggravate dimensionality issues, requiring innovative approaches to screen mined temporal patterns. Applying feature selection methods and inclusion of ontologies have been recommended to generate a small set of predictive and non-spurious temporal patterns.[6,11]

Figure 2 illustrates our study design. Adapting the framework from Moskovitch and Shahar (2015), we use a four-phase framework for classification of disease using EHR observations. We first utilize medical

ontologies to construct a set of features. Second, we apply a simple yet knowledge-driven temporal abstraction technique to construct observation sequences. In the third phase, we perform feature screening and embedded feature selection to identify a set of highly informative features, which we use in the fourth phase for the classification purpose. The experiments involve comparing classification performance obtained from raw and sequenced features at different ontological classes.

We used EHR data from a congestive heart failure (CHF) cohort. Use of data for this study was approved by the Partners Institutional Review Board (2002P000381). To create the gold standard for training, a random subset of 250 subjects was selected for full chart review by a clinical SME from the entire data mart of 177,000 subjects. Of the 250 subjects for whom charts were reviewed, 93 had CHF, 114 did not have the disease, 13 possibly had the diseases, and 30 were unknown because of too little information. The 13 with probable disease status were grouped in to the no disease category. We included the 93+114=207 patients in this study, excluding the ones with possible and indeterminate status. We further narrowed the dataset by only using the medications and diagnosis data. The 207 patients with gold standard labels for CHF had 180,821 medications and diagnoses observations.

Ontology-to-feature mapping

We utilized the i2b2 hierarchical ontology to construct the feature set for classifying Heart Failure (HF) patients. From all categories of EHR data, we used medications and diagnoses observations. Diagnoses observations were defined by International Classification of Diseases, Tenth Revision, Clinical Modification (ICD 10 CM). Using expert knowledge, we extracted three classes of features by mapping leaves from the ontology to features: (1) 4th level diagnoses and medications leaves (4D4M), 2) 3rd level diagnoses and medications leaves (3D3M), and (3) 4th level diagnoses and 3rd level medications leaves (4D3M). Observations from each of these ontology classes comprised our raw EHR features, which we employed in mining observation sequences and benchmarking classification performance obtained from using sequenced EHR data (compared with raw EHR data). The 4D4M ontology class resulted in 2,951 raw EHR features for the 207 patients, the 3D3M ontology class resulted in 393 raw EHR features, and the 4D3M ontology class resulted in 2,644 raw EHR features – by using the 3rd level leaves for medications, we did not lose much granularity, compared with the 4th level.

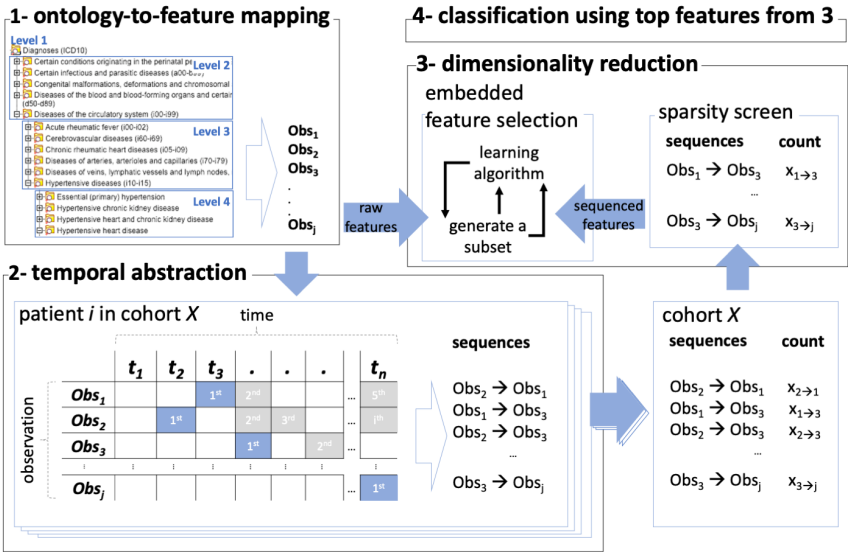


Figure 2. Study design, from feature engineering to classification

Temporal abstraction: Mining observation sequences

We only utilized the first occurrence of each unique observation at a given time, resulting in 20,001 observations when using the 4D4M ontology class, 9,142 observations when using the 3D3M, and 16,539 observations when using the 4D3M ontology classes. Using the resulting observations, we constructed 2-deep sequences (or sequenced features) by sorting observation data for each patient by date/time. A sequence pair of A→B includes observation A at time t_1 and observation B at time t_2, when t_1 is smaller than t_2 ($t_1 < t_2$). This procedure for constructing observation sequences resulted in curation of a large vector of sequences for each patient. We aggregated patient-level sequences to create a database of sequenced observations for the entire cohort.

Dimensionality reduction

Sparsity screen

Sequencing observations will produce a lot more features than the raw EHR data. for 1,000 observations, we can obtain as many as 499,500 bidirectional combinations, based on $C(n,r) = \frac{n!}{(r!(n-r)!)}$), which will result in 999,000 (499,500 × 2) unique 2-deep sequences considering directionality of sequences – i.e., both A→B and B→A are possible in a cohort. However, many of these features can be sparse that can cause dimensionality issues. To screen and address potential sparsity issues, we created frequency histograms and applied correlation analysis. We computed a matrix of Pearson's r correlation coefficients and asymptotic correlation P-values for all possible pairs of observed sequences in the cohort (count of sequences by patient) against the labeled outcome variable (whether or not the patient had HF).

Embedded feature selection

A fundamental problem in Machine Learning is to approximate the functional relationship between a feature set and an output. Feature selection is key to building robust classifiers.[74] Irrelevant features can induce greater computational cost and lead to overfitting, which would diminish the classifier's generalizability. The goal of feature selection is to find a subset of features that maximizes an objective function. Feature search methods are often characterized under filter, wrapper, and embedded procedures.[49,75–79] In embedded feature selection procedures, the classifier learns to weight features based on their contribution during the training process. Regularization methods are the most common embedded feature selection procedures. Regularized classifiers introduce additional constraint in model optimization that penalize more features in favor of model parsimony. We performed embedded feature selection, using Regularized Random Forests to identify top sequenced or raw features for each classes of ontology hierarchy.

Random Forests [80,81] are one of the most popular supervised learning methods. In addition to their relatively decent predictive performance, non-parametric attributes, handling of missing data, and ease of use, the tree regularization framework in random forests provides a straightforward and efficient solution to evaluating variable importance. In a regularized tree model, new features are successively added if they provide substantial and new predictive information about the outcome variable.[82] Decrease in a weighted impurity function or predictive accuracy are the two metrics built into the tree regularization for quantifying variable importance. We used the Gini index as the impurity function, and therefore, the Mean Decrease Gini (MDG) – a.k.a. Gini importance – for variable importance. The GINI importance measures the node purity gain by splitting a variable.[83] Permuting an important variable

results in a relatively large decrease in mean GINI gain. In an ensemble of decision trees (i.e., Random Forest), a variable's MDG is a forest-wide weighted average of the decrease in the Gini Impurity metric resulted from splitting on the variable across all of the individual trees that make up the forest. The Mean Decrease Gini provide a superior means for evaluating feature importance.[84] A higher MDG indicates higher variable importance.

We ran 30 iterations of the Regularized Random Forest classifier for each ontology class and calculated the median Mean Gini Decrease for each feature across all runs. We used bootstrap cross validation with 100 re-sampling iterations. All variables were scaled and centered. Using the median MGDs, we narrowed the list of features for training a set of classifiers to compare classification performance using raw and sequenced features across different ontology classes.

Classifier training

We trained 30 logistic regression classifiers with LASSO (L_1) regularization using bootstrap cross validation with the top features from each ontology class obtained through the feature selection process. Regularized logistic regression classifiers are the most popular classifiers for EHR phenotyping. The adaptive LASSO procedure simultaneously identifies important features and provides stable estimates of the model parameters.[85] L_1 regularization shrinks the feature space by dropping irrelevant or redundant features.[86] We used the area under the Receiver Operating Characteristic (ROC) curve AUC ROC to compare classifier performances.

RESULTS

We mined 1,013,052 unique 2-deep sequences (or sequenced features) from the 20,001 raw EHR observations extracted from the 4D4M ontology class, 78,747 from the 9,142 3D3M observations, and 738,054 from the 16,539 4D3M observations. Evidently, sequencing observations offer a lot more features, compared with the raw EHR data. On the one hand, many of these features can provide useful information in classifying the outcome variable. On the other, large number of sequenced features create dimensionality issues, given the small labeled data.

We screened the 2-deep sequenced features for sparsity and found that the sparsity issue had had at least two strands. The first was that some of the sequences were only observed just a few times. For example, we found that a large number of sequences were only observed once, even after aggregating the sequence counts at the cohort level. We removed these singular sequences as the first step towards addressing sparsity, resulting in 227,184 plural (seen more than once at the cohort level) sequenced features from the 4D4M ontology class, and 48,133 and 158,337 sequenced features from the 3D3M and 4D3M ontology classes, respectively. Nevertheless, sparsity was still present in the plural sequenced features. The top plots in Figure 3 present the two strands of the sparsity issue. The histogram on the top left demonstrates that the distribution of plural sequenced features is still considerably skewed to the left. The histogram on the top right exhibits the second strand in sparsity diagnostics. It shows that many of the sequences also belong to a very small group of patients, those who have a very large number of encounters and thus observations records.

One way to address sparsity (and hence dimensionality) in sequenced data is to apply correlation analysis to exclude sequences that do not have a significant correlation with the outcome variable. We computed a matrix of correlation coefficients (Pearson's r) and asymptotic P-values for all observed sequences against the labeled outcome variable (whether or not the patient had HF). Using the P-values, we

highlighted significance status at different levels (p<0.005, p<0.01, p<0.05). The bottom plots in Figure 3 add correlation significance to the histograms, demonstrating the proportions of sequences on the histograms that are significantly associated with Heart Failure. We found that neither of the low frequency sequences nor the ones that were sequenced from a small number of patients were not significantly correlated with the outcome variable of interest. We used the correlation significances to pre-screen features and cut the dimensionality for the embedded feature selection.

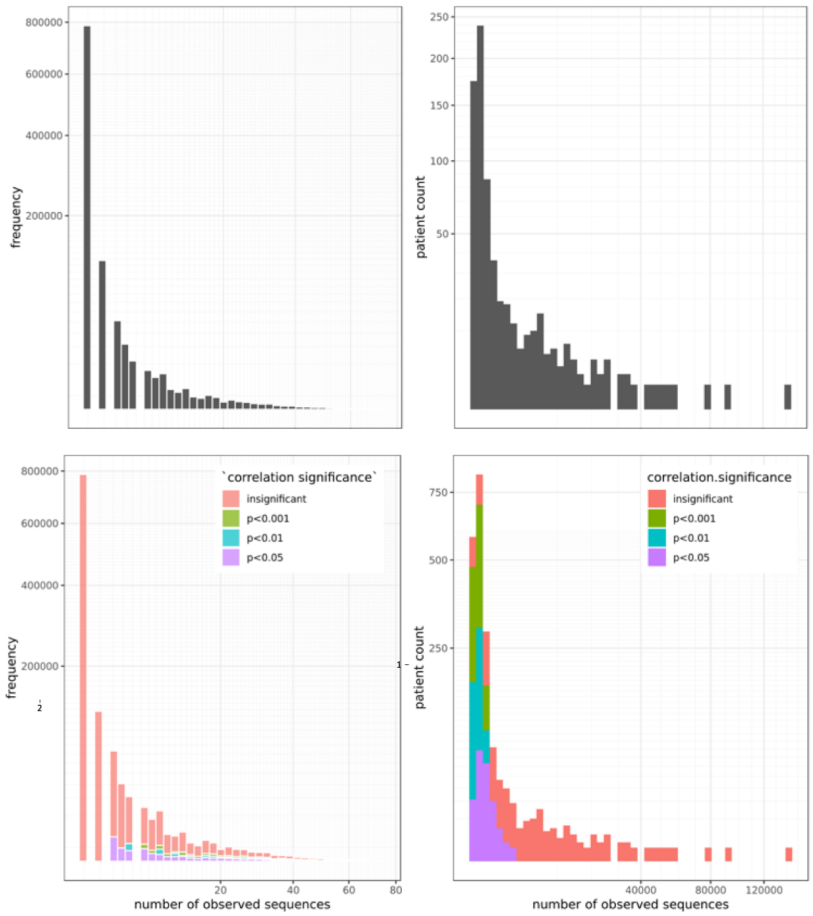


Figure 3. Correlation significance in sequence frequency and patients' contribution to sequencing

As mentioned earlier, at the 4th level of the i2b2 ontology hierarchy (the 4D4M ontology class), we obtained 227,186 non-singular observation sequences. Of the 227,186 features, 576 were correlated with the outcome variable at p<0.005, 1,141 were correlated at p<0.01, and 8,050 were correlated at p<0.05. Of the 48,135 sequences obtained from using the 3rd level of the ontology hierarchy (the 3D3M ontology class), 291 were significantly correlated at p<0.01, and 1,782 were correlated at p<0.05. Of the 158,340 sequences obtained from using the 3rd level of the ontology hierarchy for medications and 4th level for diagnoses (the 4D3M ontology class), 724 were significantly correlated at p<0.01, and 4,465 were correlated at p<0.05. We did not cut 3D3M and 4D3M sequences by p<0.005, as the number of features went to below 1,000 using sequences that were significantly correlated with the outcome variable at p<0.01. Table 1 provides a summary of features and observations obtained throughout the data processing.

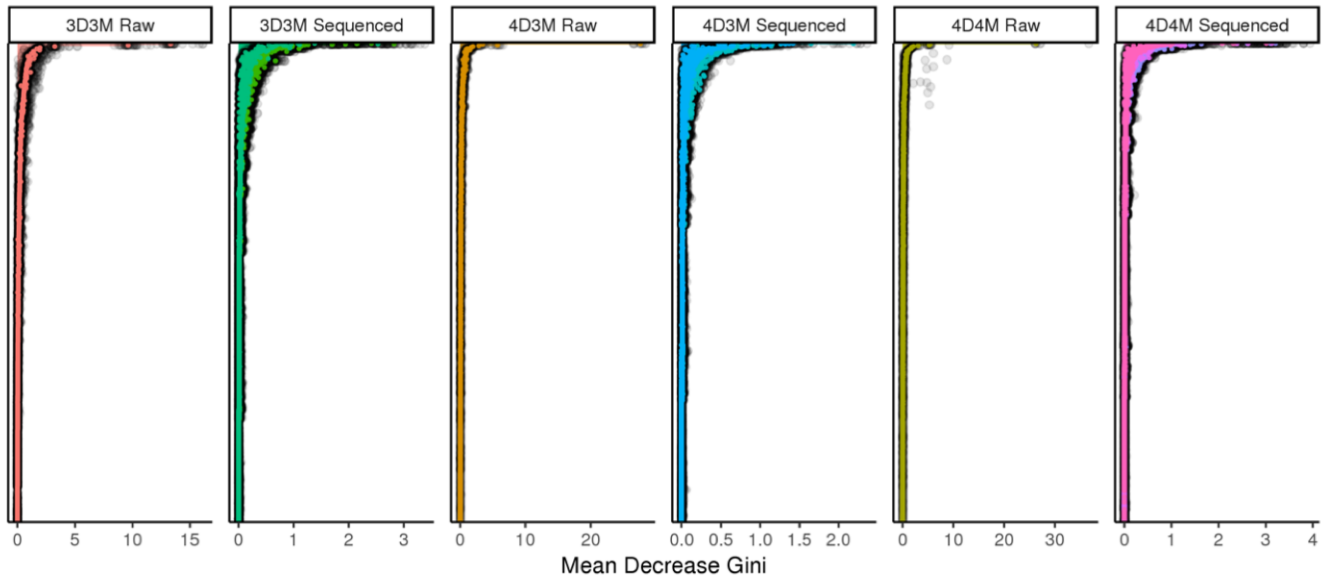
Table 1. Observation and features counts obtained from mining EHRs at different ontology levels

ontology class	raw features	first occurrence observations	2-deep sequenced features	2-deep sequenced features (plural)*	correlated features		
					0.005	*0.01*	*0.05*
4D4M**	2,951	20,001	1,013,052	227,184	576	1,141	8,050
3D3M***	393	9,142	78,747	48,133	-	291	1,782
4D3M****	2,644	16,539	738,054	158,337	-	724	4,465

*excludes singular sequences. **4th level ontology leaves for medications and diagnoses. ***3rd level ontology leaves for medications and diagnoses. ****4th level leaves for diagnoses & 3rd level leaves for medications.

Variable Importance

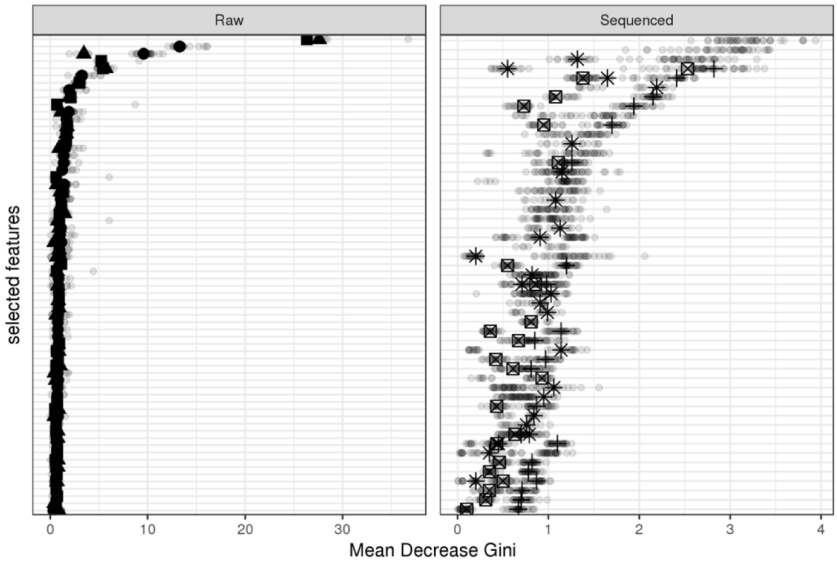
Evaluating the Mean Gini Decrease (MDG) obtained from each ontology class, we found a major difference between the raw and sequenced features. Figure 4 shows that, when using raw EHR data for classification, the classification performance depended on a very small number of features with high MDG values, whereas the remainder of features do not contribute to improving the classifiers performance. In contrast, when using the sequenced features, a larger number of features often contributed to the classification performance.



* the Y axes represent features in each ontology class, ordered by the median of Mean Decrease Gini values obtained over 30 runs. The figure shows that compared with the sequenced features, only a few raw EHR features contribute to the classification performance.

Figure 4. Ranking features by Mean Decrease Gini across the ontology class and feature type.

We further investigated this difference by zooming in the MDGs from top features by feature type (Figure 5). Top 20 features by type are also provided in Table 2. The difference in the average value of Mean Decrease Gini between the most and least important variables among the top 20 features is more than 25 for raw EHR features and less than 2 for sequenced features. This finding can be more significant if the classification performances using raw or sequenced features are comparable. That is because, in the case of comparable classification performance, classification with sequenced features would be more reliable as it would be based on a more diverse set of features.



* Dot shapes represent different ontology classes. The figure shows that the classification performance using raw EHR features relied on a very small number of features. In contrast, almost all selected sequenced features contributed to the classification, regardless of the ontology class.

Figure 5. Ranking top features by Mean Decrease Gini across feature type.

Using raw EHR observations, we found that the observation of Heart Failure in EHR data are the top features for classifying Heart Failure patients. This finding is not surprising, as 100 percent of the patients in the cohort had the diagnosis code for Heart Failure, yet only 44% were true positives. It has been shown that, due to high specificity and low sensitivity, ICD codes may be useful for ruling in cardiovascular disease, but not necessarily for ruled these disease out. [87] Further, observation of the diagnosis code in the EHR presents very limited utility for predicting Heart Failure in the future, before the onset of Heart Failure.

The sequenced features, however, present information beyond single clinical observations that can potentially shed light on some dimensions of the healthcare processes. For example, we found observation of Coumarins and indandiones after Atrial fibrillation and flutter (Atrial fibrillation and flutter → Coumarins and indandiones) is an important feature for distinguishing Heart Failure patients. This is because anticoagulation is a standard treatment to reduce stroke risk in patients with Atrial fibrillation and flutter IF THEY QUALIFY using a standard algorithm known as the CHADS score (or more recently the CHA2DS2-VASc score). The score depends on various pre-exiting conditions of the patient, one of which is Congestive Heart Failure. Therefore, one (very common) reason that patients get anticoagulation after Atrial Fibrillation is that they have Congestive Heart Failure. This implies that the two events in that order often indicate that the patient has Congestive Heart Failure, and therefore the sequence can be used to extract this reasoning from the medical record. Although only 23 percent of the patients had this sequence of observation in the medical records, 76 percent of them were true positives for Heart Failure.

Table 2. The list and average MDG values for the top 9 features at 4D4M ontology class.

	Feature	MDG *	%patients **	%positives ***
Raw	Heart failure	27.58	100%	44%
	Cardiomyopathy	5.60	33%	64%
	Loop diuretics	5.56	70%	53%
	Essential (primary) hypertension	2.94	76%	43%
	Chronic ischemic heart disease	2.02	64%	54%
	Pleural effusion in conditions classified elsewhere	1.71	52%	46%
	Nitroglycerin	1.55	43%	49%
	Abnormal findings on diagnostic imaging of lung	1.51	44%	33%
	Other respiratory disorders	1.43	61%	42%
Sequenced	Atrial fibrillation and flutter → Coumarins and indandiones	3.08	23%	76%
	Asthma → Cardio selective beta blockers	3.05	6%	100%
	Chronic ischemic heart disease → Pneumonia, unspecified organism	2.41	26%	71%
	Heparins → Salicylates	1.60	17%	69%
	Salicylates → Potassium-sparing diuretics	1.43	5%	100%
	Lisinopril → Ascites	1.36	3%	100%
	Non-cardio selective beta blockers → Lisinopril	1.19	7%	85%
	Cardiomyopathy → Encounter for preprocedural examinations	1.18	11%	90%
	Cardiomyopathy → Transplanted organ and tissue status	1.17	2%	100%

*average of the Mean Decrease Gini over 30 iterations. "→" defines the observation order in the sequence.
** percent of patient for whom feature was observed – e.g., 100 percent of patients had diagnosis code for Heart Failure. *** percent of positives from patient for whom feature was observed – e.g., of the patients who had diagnosis code for Heart Failure, only 44% had heart failure.

The reverse sequence, Anticoagulation before diagnosis of Atrial fibrillation and flutter (Coumarins and indandiones → Atrial fibrillation and flutter), would just be a sign of two totally unrelated clinical processes. For example, maybe the patient had a knee replacement and was on short term anticoagulation then much later developed Atrial fibrillation.

We also found that a sequence of cardio selective beta blockers after diagnosis of Asthma (Asthma → Cardio selective beta blockers) was an important feature for heart failure classification. Beta blockers are not typically used as treatment for Asthma. Use of beta blockers for Asthma patients is a contentious issue. As the use of beta-blockers in patients with reactive airway disease is a subject of clinical controversy, the sequence of beta-blocker following asthma diagnosis may be a marker of a change in the clinical weight of risk and benefits that reflects more serious cardiovascular disease.[88]

Do sequenced features make "better" classifiers?

From the top features, we learned that sequenced features have lower prevalence in medical records, but present more useful information for classification as the majority of those who have the sequences are likely to have the disease. To compare the performance of features by type and ontology class, we performed 30 iterations of L_1 logistic regression using the top selected features by type. We used the area under the Receiver Operating Characteristic (ROC) curve AUC ROC to compare classifier performances. Figure 6 shows that the sequenced features unanimously outperformed raw EHR features for

classification. We found that sequenced features improve classification performance from raw EHR data by more than 10 percent.

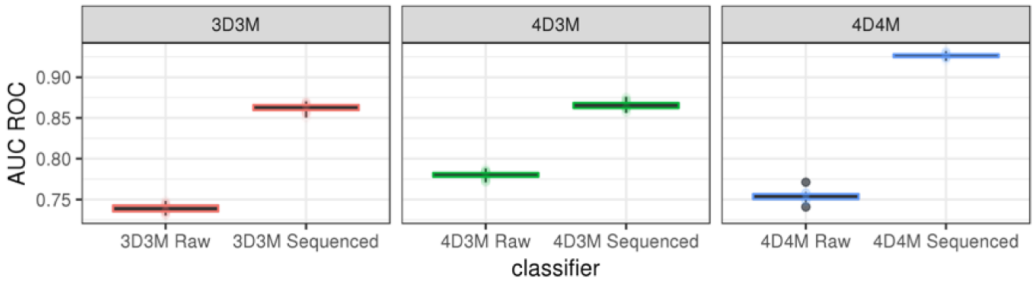


Figure 6. Comparing the area under the Receiver Operating Characteristic curve by feature type and ontology class.

Table 3. Classifier performance ranks by feature type and ontology class.

Ontology class	Feature type	Mean AUC ROC	Rank
4D4M	Raw	0.753	4
	Sequenced	0.926	1
3D3M	Raw	0.739	5
	Sequenced	0.862	2*
4D3M	Raw	0.779	3
	Sequenced	0.865	2*

*2nd place tie.

Ranking the classifiers by feature type and ontology class (Table 3), we found that using the sequenced features from the 4D4M ontology class (4th level of ontology for both diagnoses and medications) resulted in the best classifier. The difference between the 3rd and 4th level of ontology for diagnoses was not statistically significant when using sequenced features. However, using the raw EHR observations as features, when 4th level of ontology was used for diagnoses, using the 3rd level of medication ontology slightly improved the classification performance against the 4th level of medication ontology. Nevertheless, sequenced features clearly outperformed raw EHR features in classifying Heart Failure patients. Overall, we found that not only do the sequenced features carry important information about healthcare processes, they also result in significantly improving classification performance.

DISCUSSION

The conception of time relies on how we perceive it, but it can become tenseless when described with a mathematical structure.[11] Temporality of EHR observation data has not been fully exploited by current methods in biomedical research.[3] While interest in applying temporal approaches to EHR data is growing, progress is inhibited by complexities of adequately representing time in EHR observations. As a result, meaningfully changing clinical care by identifying and validating novel temporal patterns between diseases has seen limited success.[89] Efforts towards creating phenotypes and discover knowledge from EHR observations must account for the biases that are inherent in EHR data due to the recording process.[4]

Our ability to predict a patient's true health state from her medical history is integral to intervening in the progression of disease[10] and reducing the burden of healthcare.[90] We demonstrated that short

sequences of EHR observations offer a better classification performance that using singular raw observations. We also found that sequenced patterns carry more information about existence of the disease, and thus, are potentially also useful for predicting the disease onset before an observation of the disease is recorded in the EHRs. Dahlem et al. (2015) applied a similar approach, but only to disease (ICD-9 codes) codes and showed that harnessing knowledge of disease progressions increases the predictability bounds, compared to when disease histories are treated independently.[10]

Other similar work include Batal et al. (2013) who used temporal abstraction and temporal pattern mining to extract the classification features.[13] Temporal patterns can also be harnessed to represent healthcare histories, or clinical careflows,[91,92] and characterize profiles for providers and patient cohorts [93] Dagliati et al. (2017) performed temporal electronic phenotyping by sequential pattern mining and temporal data mining to extract careflows of breast cancer patients. [93]

A limitation of this work is use of International Classification of Disease codes (ICD) for constructing ontological feature classes for Heart Failure. Despite known reliability concerns and systematic variations in their use,[94] ICD codes are still frequently used in biomedical research [95–101] and can convey valuable temporal information.[102] Nevertheless, utilizing more reliable ontologies, such as PheWAS,[103,104] may result in more precise feature generation and thus improve classification performance.

In addition, we did not filter disease observations by their phenotypic expression pattern. Patients' health states evolve over diverse time scales. Acute disease like pneumonia are more isolated spontaneous occurrences, while chronic conditions such as diabetes develop and progress over years.[28] Acute conditions tend to have lower entropies, indicating of an inherent link between the predictability of disease and their phenotypic expression pattern.[102] 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.[10]

Finally, in this paper we only studied 2-deep sequences. We envisioned extracting deeper sequences. However, constructing deeper sequences would result in a much larger set of features, magnifying the sparsity issues. Using the same data, we have mined 57,672,675 3-deep sequence from observations at 4D4M ontology class (compare with the 1,013,052 2-deep sequences). Exploring predictability of deeper sequenced features will be an interesting avenue for future research.

CONCLUSION

Electronic health records contain important temporal information and present a significant opportunity to apply innovative data mining methods and discover important medical knowledge about disease progression and treatment outcomes. Data science offers tools and methodologies for high-throughput discovery of clinically meaningful data representations from the large amounts of heterogeneous longitudinal information stored in clinical data repositories. We applied a temporal high-throughput approach (including steps for temporal abstraction, dimensionality reduction, and feature importance ranking) for sequencing EHR observations to derive predictive and interpretable data representations for phenotyping from EHR data. Results demonstrated that harnessing knowledge of disease progression through temporal sequencing can improve computational disease classification (phenotyping) using EHR observations. We also found new information about clinical processes in sequenced features that are difficult to discern from raw EHR data. Given the rapidly increasing prevalence of electronic health record

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.

FUNDING STATEMENT

This work was funded through the National Human Genome Research Institute grants R01-HG009174 and U01-HG008685.

REFERENCES

[1] P.B. Jensen, L.J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care., Nat. Rev. Genet. 13 (2012) 395–405. doi:10.1038/nrg3208.

[2] J. Zhao, P. Papapetrou, L. Asker, H. Boström, Learning from heterogeneous temporal data in electronic health records, J. Biomed. Inform. 65 (2017) 105–119. doi:10.1016/j.jbi.2016.11.006.

[3] G. Hripcsak, D.J. Albers, A. Perotte, Exploiting time in electronic health record correlations, J. Am. Med. Informatics Assoc. (2011). doi:10.1136/amiajnl-2011-000463.

[4] G. Hripcsak, D.J. Albers, Next-generation phenotyping of electronic health records, J. Am. Med. Informatics Assoc. (2013). doi:10.1136/amiajnl-2012-001145.

[5] D. Agniel, I.S. Kohane, G.M. Weber, Biases in electronic health record data due to processes within the healthcare system: Retrospective observational study, BMJ. (2018). doi:10.1136/bmj.k1479.

[6] R. Moskovitch, H. Choi, G. Hripcsak, N. Tatonetti, Prognosis of Clinical Outcomes with Temporal Patterns and Experiences with One Class Feature Selection, IEEE/ACM Trans. Comput. Biol. Bioinforma. (2017). doi:10.1109/TCBB.2016.2591539.

[7] A. Singh, G. Nadkarni, O. Gottesman, S.B. Ellis, E.P. Bottinger, J. V Guttag, Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration., J. Biomed. Inform. 53 (2015) 220–8. doi:10.1016/j.jbi.2014.11.005.

[8] G. Savova, S. Bethard, W. Styler, J. Martin, M. Palmer, J. Masanz, W. Ward, Towards temporal relation discovery from the clinical narrative., AMIA ... Annu. Symp. Proceedings. AMIA Symp. (2009).

[9] R. Moskovitch, C. Walsh, G. Hripsack, N. Tatonetti, Prediction of Biomedical Events via Time Intervals Mining, 2014.

[10] D. Dahlem, D. Maniloff, C. Ratti, Predictability bounds of electronic health records, Sci. Rep. (2015). doi:10.1038/srep11865.

[11] M. Madkour, D. Benhaddou, C. Tao, Temporal data representation, normalization, extraction, and reasoning: A review from clinical domain, Comput. Methods Programs Biomed. 128 (2016) 52–68. doi:10.1016/j.cmpb.2016.02.007.

[12] R. Moskovitch, Y. Shahar, Classification-driven temporal discretization of multivariate time series, Data Min. Knowl. Discov. (2015). doi:10.1007/s10618-014-0380-z.

[13] I. Batal, H. Valizadegan, G.F. Cooper, M. Hauskrecht, A temporal pattern mining approach for

classifying electronic health record data, ACM Trans. Intell. Syst. Technol. (2013). doi:10.1145/2508037.2508044.

[14] Z. Liu, L. Wu, M. Hauskrecht, Modeling Clinical Time Series Using Gaussian Process Sequences, in: Proc. 2013 SIAM Int. Conf. Data Min., 2013. doi:10.1137/1.9781611972832.69.

[15] R. Moskovitch, F. Polubriaginof, A. Weiss, P. Ryan, N. Tatonetti, Procedure prediction from symbolic Electronic Health Records via time intervals analytics, J. Biomed. Inform. (2017). doi:10.1016/j.jbi.2017.07.018.

[16] D.J. Albers, G. Hripcsak, Estimation of time-delayed mutual information and bias for irregularly and sparsely sampled time-series, Chaos, Solitons and Fractals. (2012). doi:10.1016/j.chaos.2012.03.003.

[17] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature. (1986). doi:10.1038/323533a0.

[18] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Comput. (1997). doi:10.1162/neco.1997.9.8.1735.

[19] S.-H. Huang, C.-H. Cheng, C.-H. Chiang, C.-M. Chang, Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation Kyunghyun, Circuits Syst. 2006. APCCAS 2006. IEEE Asia Pacific Conf. (2006). doi:10.1109/APCCAS.2006.342179.

[20] E. Choi, A. Schuetz, W.F. Stewart, J. Sun, Using recurrent neural network models for early detection of heart failure onset, J. Am. Med. Informatics Assoc. (2017). doi:10.1093/jamia/ocw112.

[21] Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu, Recurrent Neural Networks for Multivariate Time Series with Missing Values, Sci. Rep. (2018). doi:10.1038/s41598-018-24271-9.

[22] Z.C. Lipton, D.C. Kale, C. Elkan, R.C. Wetzel, Learning to Diagnose with LSTM Recurrent Neural Networks, CoRR. abs/1511.0 (2015). http://arxiv.org/abs/1511.03677.

[23] T. Pham, T. Tran, D. Phung, S. Venkatesh, Predicting healthcare trajectories from medical records: A deep learning approach, J. Biomed. Inform. (2017). doi:10.1016/j.jbi.2017.04.001.

[24] B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin, X. Wei, Predicting the Risk of Heart Failure With EHR Sequential Data Modeling, IEEE Access. 6 (2018) 9256–9261. doi:10.1109/ACCESS.2017.2789324.

[25] A.N. Jagannatha, H. Yu, Bidirectional RNN for Medical Event Detection in Electronic Health Records., Proc. Conf. Assoc. Comput. Linguist. North Am. Chapter. Meet. 2016 (2016) 473–482.

[26] E. Choi, M.T. Bahadori, A. Schuetz, W.F. Stewart, J.S.B.T.-P. of the 1st M.L. for H. Conference, Doctor AI: Predicting Clinical Events via Recurrent Neural Networks, (2016) 301–318. http://proceedings.mlr.press/v56/Choi16.pdf.

[27] J. Zhang, K. Kowsari, J.H. Harrison, J.M. Lobo, L.E. Barnes, Patient2Vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record, IEEE Access. (2018). doi:10.1109/ACCESS.2018.2875677.

[28] R. Pivovarov, N. Elhadad, Automated methods for the summarization of electronic health records, J. Am. Med. Informatics Assoc. (2015). doi:10.1093/jamia/ocv032.

[29] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records, Sci. Rep. 6 (2016). doi:10.1038/srep26094.

[30] S. Yu, Y. Ma, J. Gronsbell, T. Cai, A.N. Ananthakrishnan, V.S. Gainer, S.E. Churchill, P. Szolovits, S.N. Murphy, I.S. Kohane, K.P. Liao, T. Cai, Enabling phenotypic big data with PheNorm, J. Am. Med. Informatics Assoc. 25 (2018) 54–60. doi:10.1093/jamia/ocx111.

[31] V. Agarwal, T. Podchiyska, J.M. Banda, V. Goel, T.I. Leung, E.P. Minty, T.E. Sweeney, E. Gyang, N.H. Shah, Learning statistical models of phenotypes using noisy labeled training data, J. Am. Med. Informatics Assoc. 23 (2016) 1166–1173. doi:10.1093/jamia/ocw028.

[32] S. Yu, A. Chakrabortty, K.P. Liao, T. Cai, A.N. Ananthakrishnan, V.S. Gainer, S.E. Churchill, P. Szolovits, S.N. Murphy, I.S. Kohane, T. Cai, Surrogate-assisted feature extraction for high-throughput phenotyping, J. Am. Med. Inform. Assoc. 24 (2017) e143–e149. doi:10.1093/jamia/ocw135.

[33] S. Yu, K.P. Liao, S.Y. Shaw, V.S. Gainer, S.E. Churchill, P. Szolovits, S.N. Murphy, I.S. Kohane, T. Cai, Toward high-throughput phenotyping: Unbiased automated feature extraction and selection from knowledge sources, J. Am. Med. Informatics Assoc. 22 (2015) 993–1000. doi:10.1093/jamia/ocv034.

[34] K.P. Liao, T. Cai, V. Gainer, S. Goryachev, Q. Zeng-treitler, S. Raychaudhuri, P. Szolovits, S. Churchill, S. Murphy, I. Kohane, E.W. Karlson, R.M. Plenge, Electronic medical records for discovery research in rheumatoid arthritis., Arthritis Care Res. (Hoboken). (2010). doi:10.1002/acr.20184.

[35] A.N. Ananthakrishnan, T. Cai, G. Savova, S.-C. Cheng, P. Chen, R.G. Perez, V.S. Gainer, S.N. Murphy, P. Szolovits, Z. Xia, S. Shaw, S. Churchill, E.W. Karlson, I. Kohane, R.M. Plenge, K.P. Liao, Improving Case Definition of Crohn's Disease and Ulcerative Colitis in Electronic Medical Records Using Natural Language Processing: A Novel Informatics Approach., Inflamm. Bowel Dis. (2013). doi:10.1097/MIB.0b013e31828133fd.

[36] Z. Xia, E. Secor, L.B. Chibnik, R.M. Bove, S. Cheng, T. Chitnis, A. Cagan, V.S. Gainer, P.J. Chen, K.P. Liao, S.Y. Shaw, A.N. Ananthakrishnan, P. Szolovits, H.L. Weiner, E.W. Karlson, S.N. Murphy, G.K. Savova, T. Cai, S.E. Churchill, R.M. Plenge, I.S. Kohane, P.L. De Jager, Modeling disease severity in multiple sclerosis using electronic health records, PLoS One. (2013). doi:10.1371/journal.pone.0078927.

[37] V. Kumar, K. Liao, S.-C. Cheng, S. Yu, U. Kartoun, A. Brettman, V. Gainer, A. Cagan, S. Murphy, G. Savova, P. Chen, P. Szolovits, Z. Xia, E. Karlson, R. Plenge, A. Ananthakrishnan, S. Churchill, T. Cai, I. Kohane, S. Shaw, Natural language processing improves phenotypic accuracy in an electronic medical record cohort of type 2 diabetes and cardiovascular disease, J. Am. Coll. Cardiol. (2014). doi:10.1016/S0735-1097(14)61359-0.

[38] S.N. Murphy, M.E. Mendis, D.A. Berkowitz, I. Kohane, H.C. Chueh, Integration of clinical and genetic data in the i2b2 architecture, AMIA Annu. Symp. Proc. (2006) 1040. doi:85881 [pii].

[39] V.M. Castro, J. Minnier, S.N. Murphy, I. Kohane, S.E. Churchill, V. Gainer, T. Cai, A.G. Hoffnagle, Y. Dai, S. Block, S.R. Weill, M. Nadal-Vicens, A.R. Pollastri, J.N. Rosenquist, S. Goryachev, D. Ongur, P. Sklar, R.H. Perlis, J.W. Smoller, P.H. Lee, E.A. Stahl, S.M. Purcell, D.M. Ruderfer, A.W. Charney, P. Roussos, C. Pato, M. Pato, H. Medeiros, J. Sobel, N. Craddock, I. Jones, L. Forty, A. DiFlorio, E. Green, L. Jones, K. Dunjewski, M. Landén, C. Hultman, A. Juréus, S. Bergen, O. Svantesson, S. McCarroll, J.

Moran, K. Chambert, R.A. Belliveau, Validation of electronic health record phenotyping of bipolar disorder cases and controls, Am. J. Psychiatry. (2015). doi:10.1176/appi.ajp.2014.14030423.

[40] S. Yu, K.K. Kumamaru, E. George, R.M. Dunne, A. Bedayat, M. Neykov, A.R. Hunsaker, K.E. Dill, T. Cai, F.J. Rybicki, Classification of CT pulmonary angiography reports by presence, chronicity, and location of pulmonary embolism with natural language processing, J. Biomed. Inform. (2014). doi:10.1016/j.jbi.2014.08.001.

[41] K.P. Liao, A.N. Ananthakrishnan, V. Kumar, Z. Xia, A. Cagan, V.S. Gainer, S. Goryachev, P. Chen, G.K. Savova, D. Agniel, S. Churchill, J. Lee, S.N. Murphy, R.M. Plenge, P. Szolovits, I. Kohane, S.Y. Shaw, E.W. Karlson, T. Cai, Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts, PLoS One. (2015). doi:10.1371/journal.pone.0136651.

[42] K.P. Liao, T. Cai, G.K. Savova, S.N. Murphy, E.W. Karlson, A.N. Ananthakrishnan, V.S. Gainer, S.Y. Shaw, Z. Xia, P. Szolovits, S. Churchill, I. Kohane, Development of phenotype algorithms using electronic medical records and incorporating natural language processing., Br. Med. J. (2015). doi:10.1136/bmj.h1885.

[43] D.M. Roden, J.M. Pulley, M.A. Basford, G.R. Bernard, E.W. Clayton, J.R. Balser, D.R. Masys, Development of a large-scale de-identified DNA biobank to enable personalized medicine, Clin. Pharmacol. Ther. (2008). doi:10.1038/clpt.2008.89.

[44] E.A. Bowton, S.P. Collier, X. Wang, C.B. Sutcliffe, S.L. Van Driest, L.J. Couch, M. Herrera, R.N. Jerome, R.J.C. Slebos, W.E. Alborn, D.C. Liebler, C.D. McNaughton, R.L. Mernaugh, Q.S. Wells, N.J. Brown, D.M. Roden, J.M. Pulley, Phenotype-Driven Plasma biobanking strategies and Methods, J. Pers. Med. (2015). doi:10.3390/jpm5020140.

[45] C.A. McCarty, R.L. Chisholm, C.G. Chute, I.J. Kullo, G.P. Jarvik, E.B. Larson, R. Li, D.R. Masys, M.D. Ritchie, D.M. Roden, J.P. Struewing, W.A. Wolf, The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies, BMC Med. Genomics. 4 (2011) 13. doi:10.1186/1755-8794-4-13.

[46] O. Gottesman, H. Kuivaniemi, G. Tromp, W.A. Faucett, R. Li, T.A. Manolio, S.C. Sanderson, J. Kannry, R. Zinberg, M.A. Basford, M. Brilliant, D.J. Carey, R.L. Chisholm, C.G. Chute, J.J. Connolly, D. Crosslin, J.C. Denny, C.J. Gallego, J.L. Haines, H. Hakonarson, J. Harley, G.P. Jarvik, I. Kohane, I.J. Kullo, E.B. Larson, C. McCarty, M.D. Ritchie, D.M. Roden, M.E. Smith, E.P. Böttinger, M.S. Williams, The Electronic Medical Records and Genomics (eMERGE) Network: Past, present, and future, Genet. Med. (2013). doi:10.1038/gim.2013.72.

[47] J. Pathak, J. Wang, S. Kashyap, M. Basford, R. Li, D.R. Masys, C.G. Chute, Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience., J. Am. Med. Inform. Assoc. 18 (n.d.) 376–86. doi:10.1136/amiajnl-2010-000061.

[48] A.N. Kho, J.A. Pacheco, P.L. Peissig, L. Rasmussen, K.M. Newton, N. Weston, P.K. Crane, J. Pathak, C.G. Chute, S.J. Bielinski, I.J. Kullo, R. Li, T.A. Manolio, R.L. Chisholm, J.C. Denny, Electronic medical records for genetic research: Results of the eMERGE consortium, Sci. Transl. Med. (2011). doi:10.1126/scitranslmed.3001807.

[49] I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection, J. Mach. Learn. Res. 3 (2003) 1157–1182. doi:10.1016/j.aca.2011.07.027.

[50] R. Moskovitch, Y. Shahar, Medical Temporal-Knowledge Discovery via Temporal Abstraction, in: AMIA Annu. Symp. Proc., 2009. doi:10.1016/j.cryogenics.2004.07.006.

[51] Y. Shahar, A Framework for Knowledge-Based Temporal Abstraction, Artif. Intell. (1997). doi:10.1016/S0004-3702(96)00025-2.

[52] M. Stacey, C. McGregor, Temporal abstraction in intelligent clinical data analysis: A survey, Artif. Intell. Med. (2007). doi:10.1016/j.artmed.2006.08.002.

[53] L. Sacchi, A. Dagliati, R. Bellazzi, Analyzing complex patients' temporal histories: New frontiers in temporal data mining, Methods Mol. Biol. (2015). doi:10.1007/978-1-4939-1985-7_6.

[54] M.J. O'Connor, W.E. Grosso, S.W. Tu, M.A. Musen, RASTA: A distributed temporal abstraction system to facilitate knowledge-driven monitoring of clinical databases, in: Stud. Health Technol. Inform., 2001. doi:10.3233/978-1-60750-928-8-508.

[55] M. Monroe, B. Shneiderman, C. Plaisant, J. Morales, J. Millstein, The Challenges of Specifying Intervals and Absences in Temporal Queries : A Graphical Language Approach, ACM CHI Conf. Hum. Factors Comput. Syst. (2013). doi:10.1145/2470654.2481325.

[56] W. Horn, S. Miksch, G. Egghart, C. Popow, F. Paky, Effective data validation of high-frequency data: Time-point-, time- interval-, and trend-based methods, Comput. Biol. Med. (1997). doi:10.1016/S0010-4825(97)00012-7.

[57] Y. Shahar, C. Cheng, Model-based visualization of temporal abstractions, Comput. Intell. (2000). doi:10.1111/0824-7935.00114.

[58] A.S. Silvent, M. Dojat, C. Garbay, Multi-level temporal abstraction for medical scenario construction, Int. J. Adapt. Control Signal Process. (2005). doi:10.1002/acs.855.

[59] Y. Shahar, M.A. Musen, Knowledge-based temporal abstraction in clinical domains, Artif. Intell. Med. (1996). doi:10.1016/0933-3657(95)00036-4.

[60] A. Seyfang, S. Miksch, Advanced temporal data abstraction for guideline execution, in: Stud. Health Technol. Inform., 2004. doi:10.3233/978-1-60750-944-8-88.

[61] J. Hunter, N. McIntosh, Knowledge-Based Event Detection in Complex Time Series Data, Proc. Jt. Eur. Conf. Artif. Intell. Med. Med. Decis. Mak. (1999). doi:10.1007/3-540-48720-4.

[62] G. Carrault, M.O. Cordier, R. Quiniou, F. Wang, Temporal abstraction and inductive logic programming for arrhythmia recognition from electrocardiograms, Artif. Intell. Med. (2003). doi:10.1016/S0933-3657(03)00066-6.

[63] K. Orphanou, A. Stassopoulou, E. Keravnou, Temporal abstraction and temporal Bayesian networks in clinical domains: A survey, Artif. Intell. Med. (2014). doi:10.1016/j.artmed.2013.12.007.

[64] G. Hripcsak, N. Elhadad, Y.H. Chen, L. Zhou, F.P. Morrison, Using Empiric Semantic Correlation to Interpret Temporal Assertions in Clinical Texts, J. Am. Med. Informatics Assoc. (2009). doi:10.1197/jamia.M3007.

[65] Y. Xu, Y. Wang, T. Liu, J. Tsujii, E.I.C. Chang, An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge, J. Am. Med. Informatics Assoc. (2013). doi:10.1136/amiajnl-2012-001607.

[66] N.Y. Jane, K.H. Nehemiah, K. Arputharaj, A Temporal Mining Framework for Classifying Un-Evenly Spaced Clinical Data An Approach for Building Effective Clinical Decision-Making System, Appl. Clin. Inform. (2016). doi:10.4338/ACI-2015-08-RA-0102.

[67] L.C. Cheng, Y.H. Hu, S.H. Chiou, Applying the Temporal Abstraction Technique to the Prediction of Chronic Kidney Disease Progression, J. Med. Syst. (2017). doi:10.1007/s10916-017-0732-5.

[68] J.H. Chen, T. Podchiyska, R.B. Altman, OrderRex: Clinical order decision support and outcome predictions by data-mining electronic medical records, J. Am. Med. Informatics Assoc. (2016). doi:10.1093/jamia/ocv091.

[69] F. Khoshnevisan, J. Ivy, M. Capan, R. Arnold, J. Huddleston, M. Chi, Recent temporal pattern mining for septic shock early prediction, in: Proc. - 2018 IEEE Int. Conf. Healthc. Informatics, ICHI 2018, 2018. doi:10.1109/ICHI.2018.00033.

[70] A.M. Cole, K.A. Stephens, G.A. Keppel, H. Estiri, L.-M. Baldwin, Extracting Electronic Health Record Data in a Practice-Based Research Network: Processes to Support Translational Research across Diverse Practice Organizations., EGEMS (Washington, DC). 4 (2016) 1206. doi:10.13063/2327-9214.1206.

[71] A.R. Post, T. Kurc, R. Willard, H. Rathod, M. Mansour, A.K. Pai, W.M. Torian, S. Agravat, S. Sturm, J.H. Saltz, Temporal abstraction-based clinical phenotyping with Eureka!, AMIA Annu Symp Proc. (2013).

[72] M. Vidal, M.E. Cusick, A.L. Barabási, Interactome networks and human disease, Cell. (2011). doi:10.1016/j.cell.2011.02.016.

[73] D. Fradkin, F. Mörchen, Mining sequential patterns for classification, Knowl. Inf. Syst. (2015). doi:10.1007/s10115-014-0817-0.

[74] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2009. doi:10.1007/b94608.

[75] L. Ladha, T. Deepa, Feature selection methods and algorithms, Int. J. …. 3 (2011) 1787–1797. http://journals.indexcopernicus.com/abstract.php?icid=945099.

[76] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics. 23 (2007) 2507–2517. doi:10.1093/bioinformatics/btm344.

[77] M. Dash, H. Liu, L. H., Feature selection for classification, Intell. Data Anal. 1 (1997) 131–156. doi:10.3233/ida-1997-1302.

[78] L. Yu, H. Liu, Efficient Feature Selection via Analysis of Relevance and Redundancy, J. Mach. Learn. Res. 5 (2004) 1205–1224. doi:10.1145/1014052.1014149.

[79] H. Liu, L. Yu, S.S. Member, L. Yu, S.S. Member, Toward integrating feature selection algorithms for classification and clustering, Knowl. Data Eng. IEEE Trans. 17 (2005) 491–502. doi:10.1109/TKDE.2005.66.

[80] L. Breiman, Random Forrest, Mach. Learn. (2001). doi:10.1023/A:1010933404324.

[81] L. Breiman, A. Cutler, Setting up, using, and understanding random forests V4. 0, Univ. California, Dep. Stat. (2003). doi:10.1017/CBO9781107415324.004.

[82] H. Deng, G. Runger, Feature selection via regularized trees, in: Proc. Int. Jt. Conf. Neural Networks, 2012. doi:10.1109/IJCNN.2012.6252640.

[83] G. Louppe, L. Wehenkel, A. Sutera, P. Geurts, Understanding variable importances in forests of randomized trees, Neural Inf. Process. Syst. (2013). doi:NIPS2013_4928.

[84] B.H. Menze, B.M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, F.A. Hamprecht, A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data, BMC Bioinformatics. (2009). doi:10.1186/1471-2105-10-213.

[85] H. Zou, The adaptive lasso and its oracle properties, J. Am. Stat. Assoc. (2006). doi:10.1198/016214506000000735.

[86] R. Tibshirani, Regression shrinkage and selection via the lasso: A retrospective, J. R. Stat. Soc. Ser. B Stat. Methodol. (2011). doi:10.1111/j.1467-9868.2011.00771.x.

[87] E. Birman-Deych, A.D. Waterman, Y. Yan, D.S. Nilasena, M.J. Radford, B.F. Gage, Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors, Med. Care. (2005). doi:10.1097/01.mlr.0000160417.39497.a9.

[88] D.R. Morales, C. Jackson, B.J. Lipworth, P.T. Donnan, B. Guthrie, Adverse respiratory effect of acute β -Blocker exposure in asthma: A systematic review and meta-analysis of randomized controlled trials, Chest. (2014). doi:10.1378/chest.13-1235.

[89] H. Paik, M.J. Kan, N. Rappoport, D. Hadley, M. Sirota, B. Chen, U. Manber, S.B. Cho, A.J. Butte, Tracing diagnosis trajectories over millions of inpatients reveal an unexpected association between schizophrenia and rhabdomyolysis, BioRxiv. (2018) 473082. doi:10.1101/473082.

[90] S. Reardon, A world of chronic disease, Science (80-.). (2011). doi:10.1126/science.333.6042.558.

[91] S. Quaglini, M. Stefanelli, G. Lanzola, V. Caporusso, S. Panzarasa, Flexible guideline-based patient careflow systems, Artif. Intell. Med. (2001). doi:10.1016/S0933-3657(00)00100-7.

[92] S. Quaglini, M. Stefanelli, A. Cavallini, G. Micieli, C. Fassino, C. Mossa, Guideline-based careflow systems, Artif. Intell. Med. (2000). doi:10.1016/S0933-3657(00)00050-6.

[93] A. Dagliati, L. Sacchi, A. Zambelli, V. Tibollo, L. Pavesi, J.H. Holmes, R. Bellazzi, Temporal electronic phenotyping by mining careflows of breast cancer patients, J. Biomed. Inform. (2017). doi:10.1016/j.jbi.2016.12.012.

[94] T.E. Chang, J.H. Lichtman, L.B. Goldstein, M.G. George, Accuracy of ICD-9-CM codes by hospital characteristics and stroke severity: Paul coverdell national acute stroke program, J. Am. Heart Assoc. (2016). doi:10.1161/JAHA.115.003056.

[95] D. Aronsky, P.J. Haug, C. Lagor, N.C. Dean, Accuracy of administrative data for identifying patients with pneumonia, Am. J. Med. Qual. (2005). doi:10.1177/1062860605280358.

[96] L.B. Goldstein, Accuracy of ICD-9-CM coding for the identification of patients with acute ischemic stroke: Effect of modifier codes, Stroke. (1998). doi:10.1161/01.STR.29.8.1602.

[97] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.C. Luthi, L.D. Saunders, C.A. Beck, T.E. Feasby, W.A. Ghali, Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data, Med. Care. (2005). doi:10.1097/01.mlr.0000182534.19832.83.

[98] G. Van Belleghem, S. Devos, L. De Wit, I. Hubloue, D. Lauwaert, K. Pien, K. Putman, Predicting in-hospital mortality of traffic victims: A comparison between AIS-and ICD-9-CM-related injury severity scales when only ICD-9-CM is reported, Injury. (2016). doi:10.1016/j.injury.2015.08.025.

[99] M. Popescu, M. Khalilia, Improving disease prediction using ICD-9 ontological features, in: IEEE Int. Conf. Fuzzy Syst., 2011. doi:10.1109/FUZZY.2011.6007410.

[100] L.I. Lezzoni, S.M. Foley, J. Daley, J. Hughes, E.S. Fisher, T. Heeren, Comorbidities, Complications, and Coding Bias: Does the Number of Diagnosis Codes Matter in Predicting In-Hospital Mortality?, JAMA J. Am. Med. Assoc. (1992). doi:10.1001/jama.1992.03480160055034.

[101] D.C. Hsia, W.M. Krushat, A.B. Fagan, J.A. Tebbutt, R.P. Kusserow, Accuracy of Diagnostic Coding for Medicare Patients under the Prospective-Payment System, N. Engl. J. Med. (1988). doi:10.1056/NEJM198802113180604.

[102] A. Perotte, G. Hripcsak, Temporal properties of diagnosis code time series in aggregate, IEEE J. Biomed. Heal. Informatics. (2013). doi:10.1109/JBHI.2013.2244610.

[103] J.C. Denny, M.D. Ritchie, M.A. Basford, J.M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D.R. Masys, D.M. Roden, D.C. Crawford, PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations, Bioinformatics. (2010). doi:10.1093/bioinformatics/btq126.

[104] S.A. Pendergrass, K. Brown-Gentry, S.M. Dudek, E.S. Torstenson, J.L. Ambite, C.L. Avery, S. Buyske, C. Cai, M.D. Fesinmeyer, C. Haiman, G. Heiss, L.A. Hindorff, C.N. Hsu, R.D. Jackson, C. Kooperberg, L. Le Marchand, Y. Lin, T.C. Matise, L. Moreland, K. Monroe, A.P. Reiner, R. Wallace, L.R. Wilkens, D.C. Crawford, M.D. Ritchie, The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery, Genet. Epidemiol. (2011). doi:10.1002/gepi.20589.