PATTERNS IN BIOMEDICAL DATA-HOW DO WE FIND THEM?

ANNA O. BASILE
The Pennsylvania State University, Department of Biochemistry and Molecular Biology
328 Innovation Blvd Ste 210
State College, PA 16803
azo121@psu.edu

ANURAG VERMA
Geisinger Health System
The Pennsylvania State University, Huck Institutes of the Life Sciences
328 Innovation Blvd Ste 210
State College, PA 16803
averma@geisinger.edu

MARTA BYRSKA-BISHOP
Geisinger Health System
328 Innovation Blvd Ste 210
State College, PA 16803
mbyrskabishop@geisinger.edu

SARAH A. PENDERGRASS
Geisinger Health System, Biomedical and Translational Informatics
122 Weis Center for Research
Danville, PA 17822
spendergrass@geisinger.edu

CHRISTIAN DARABOS
Dartmouth College, Research Computing Services
HB 6129
Hanover, NH 03755
christian.darabos@dartmouth.edu

H. LESTER KIRCHNER
Geisinger Health System, Biomedical and Translational Informatics
100 N. Academy Ave
Danville, PA 17822-4400
hlkirchner@geisinger.edu

Given the exponential growth of biomedical data, researchers are faced with numerous challenges in extracting and interpreting information from these large, high-dimensional, incomplete, and often noisy data. To facilitate addressing this growing concern, the "Patterns in Biomedical Data-How do we find them?" session of the 2017 Pacific Symposium on Biocomputing (PSB) is devoted to exploring pattern recognition using data-driven approaches for biomedical and precision medicine applications. The papers selected for this session focus on novel machine learning techniques as well as applications of established methods to heterogeneous data. We also feature manuscripts aimed at addressing the current challenges associated with the analysis of biomedical data.

1. Introduction

With great technological advances and numerous ‘big data’ initiatives targeted at generating and acquiring large amounts of biomedical information, there has been an astonishing growth in the volume of data in recent years [1]. Considering sequencing data alone, the size of data has approximately doubled every six months in the last decade [2]. Continuing at this rate, we can expect to reach a zettabyte of sequencing data generated per year by 2025 [2].

Thus, the age of big data is upon us, and with its arrival comes the potential to revolutionize many aspects of our lives. Decisions previously made using carefully constructed, simulated models of reality can now be made using measured data. While the term ‘big data’ is not well defined, it will be used herein to describe a situation where the amount of information far exceeds that which has been previously available [3]. Big data analyses impact many areas of society, culture, and research. To combat crime, law enforcement officials are employing seismology-like models to predict areas of high crime, and intervene to prevent them from occurring [3]. With large scale surveys, such as the Two Micron All-Sky Survey, which contains a petabyte of data, astronomers can now focus their efforts illuminating structures and exploring potential connections and hypotheses [4]. In the area of public health and precision medicine, large-scale efforts have been made to create datasets aimed at elucidating the genetic underpinnings of various traits as a means of disease prevention and development of effective treatment. For example, the Precision Medicine Initiative Cohort Program announced by President Obama plans to enroll one million participants spanning a multitude of age and race groups within the US [5]. Other large-scale genome projects include the UK 100,000 Genomes Project [6], and the Geisinger MyCode Community Health Initiative which unites Geisinger Health System and Regeneron Genetics Center in a collaboration aimed at bio-banking and whole-exome sequencing more than 200,000 patients [7]. Likewise, public datasets, such as The Cancer Genome Atlas (TCGA), which provides molecular characterization of cancer genomes, continue to provide a wealth of data to researchers with the hope of one day improving clinical patient care.

While these potentials are truly revolutionary, there are a number of challenges that can impede the promises of big data and make it difficult to extract the true value of this information. The sheer volume of available data and the rate at which it is being generated is overwhelming the majority of industries, many of which do not yet have the proper management, storage and analytical means of assessing this information [8]. Additionally, while small sample sizes are often prohibitive in research, the large sample sizes provided by big data initiatives may not be a panacea. Large sample sizes may be of little value if they are not representative of the population being assessed, are missing information (especially if missingness is nonrandom or important data is completely missing), or contain sampling biases [9]. Machine-learning approaches in this data-driven space will require an integration of different generated data types. In a biomedical setting, this may include clinical measurements, drug usage data, mRNA expression levels, and environmental exposures. These informatics methods must also be robust to incompleteness and variable sparsity, as well as heterogeneity which can present mixtures of categorical and numerical data. Further considerations that will need to be made include scalability and dealing with a feature space that far exceeds the number of samples.

The collection of papers presented in this session demonstrates a diversity of data-driven, pattern recognition approaches and challenges within the biomedical and precision health setting. These manuscripts span a wide range of categories from applications of well-studied informatics methods to novel pattern recognition techniques as well as approaches of overcoming big data challenges.

2. Session Contributions:

2.1 Machine Learning and Deep Learning Approaches

Machine learning and deep learning have recently received a great deal of attention due to their potentially transformative applications to big data. Machine learning refers to a class of algorithms that can learn from and also make predictions on data [10], while deep learning describes a branch of machine learning that models data using multiple levels of representation and abstraction. These methods do not require explicit rules as they rely on the data, and generally speaking, the more data, the better the outcome of these techniques. While the use of data-driven approaches is not new, this is an expanding area of biomedical research that is gaining momentum due to algorithmic sophistication, computational advancement, and the growth in volume and variety of available data.

Shameer et al. describe a data-driven feature selection and machine learning approach to predict hospital readmission in heart failure (HF) patients from electronic health records (EHR) in "*Predictive Modeling of Hospital Readmission Rates using Electronic Medical Record-wide Machine Learning: A Cased-Study Using Mount Sinai Health Cohort".* Several data domains were extracted from the EHR including diagnoses, medications, laboratory measurements, procedures, and vitals. Separate models were generated from the data domains using the Naïve Bayes algorithm and then combined. Feature selection was performed using a correlation-based method. Their approach was contrasted to using logistic regression, and it performed well over all existing predictive models in HF.

In the manuscript "*Missing data imputation in the electronic health record using deeply learned autoencoders*" **Beaulieu-Jones** et al. tackle the important issue of dealing with missing data, commonly encountered in the context of EHR. Specifically, the authors use the Pooled Resource Open-Access Amyotrophic Lateral Sclerosis (ALS) Clinical Trial Database (PRO-ACT) to evaluate missing data imputation performance of a machine learning approach, namely deeply learned autoencoders, and compare it to the performance of several established imputation strategies, such as mean, median, K-nearest neighbors, or Singular Value Decomposition (SVD). They show that autoencoders outperform other methods in imputation of data missing completely at random (MCAR), as well as data missing not at random (MNAR). Furthermore, they used data imputed by different methods to predict ALS progression and identify the most important predictors of ALS.

One of the challenges associated with applying machine learning approaches to biological problems is the interpretation of the models that arise from them. In the manuscript titled "*DG-Dashboard: Visualizing and Understanding Genomic Sequences Using Deep Neural Networks*", **Lanchantin** et al. present a visualization toolkit called the Deep Genomic Dashboard (DG-Dashboard), which facilitates interpretation of deep neural network models in the context of predicting transcription factor binding sites (TFBS) along genomic DNA. In particular, DG-Dashboard offers three strategies: saliency maps, temporal output scores, and class optimizations, which enable visualization of nucleotide importance within a particular motif, critical positions along a DNA sequence, as well as class-specific motif patterns for a particular TF based on predictions obtained from convolutional neural networks (CNNs), recurrent neural networks (RNNs), as well as convolutional-recurrent neural networks (CNN-RNNs). In addition to facilitating interpretation of the three deep neural network architectures, Lanchantin et al. demonstrate that CNN-RNNs outperform CNN and RNN in classification of TFBSs.

2.2 Pattern recognition applications in EHR, Medical Imaging, and Mobile Health data

Applications of machine learning approaches are widespread in the biomedical sector. EHRs, biomedical images, and mobile health apps are just a few of the many sources researchers are mining to advance human health. Data-driven approaches can leverage the wealth of information in these sources and extract meaningful knowledge which can then be utilized to study disease progression and symptom patterns, classify patient subgroups, and inform clinical practice and decision-making.

One such application is digital image analysis that was implemented to classify the bone cancer in “*Large Scale Image Segmentation and Classification for Viable and non-viable Tumor Identification in Osteosarcoma*”. **Arunachalam** et al. demonstrate a high-throughput approach to classify the tumor region from images of Hematoxylin and eosin (H&E) stain slides from bone cancer patients. They proposed a multi-tier approach where they used pixel and object based approach to color and classify different histopathological regions of cancer cells in the digital stain images. Further, they used a combination of multiple clustering algorithms to define viable and non-viable tumors.

In “*Development and Performance of Text-Mining Algorithms to Extract Socioeconomics Status from DE-identified Electronic Health Records*”, **Hollister** et al. describe a data mining approach, where they developed an algorithm to define a phenotype status from variety of structured and unstructured free text in EHR. In order to investigate socioeconomics status (SES) they developed seven different algorithms predictive of SES like Education, Occupation, Insurance Status, Retirement, Medicaid, and Homelessness. Their work addresses an important question associated with health outcomes and the socioeconomic status extracted from various semantic categories. They provide performance metric of seven algorithms, but also highlight many shortcoming and challenges that potentially affect phenotype algorithm development in current EHR systems.

In “*Methods for Clustering Time Series Data Acquired from Mobile Health Apps*”, **Tignor** and colleagues present a method to cluster individuals with asthma using data collected from a mobile health app. The data represent a time series of daily asthma symptoms which exhibit non-ignorable missingness. Their work focuses on developing a novel probabilistic imputation method, and combined with a consensus clustering algorithm, is used to identify distinct symptom patterns. Variations on the algorithm implementation are devised and compared.

Studying the heterogeneous patterns of disease manifestation and progression is important for the clinical treatment and management of a condition. In “*Learning Attributes of Disease Progression from Trajectories of Sparse Lab Values*”, **Agarwal** et al. use the Functional Clustering Model (FCM) to cluster sparse clinical lab measures from patients with Chronic Kidney Disease (CKD) from the Stanford Health Care (SHC) system. The authors hypothesize that using data-driven approaches on trajectories of sparse lab values can create clinically meaningful clusters that highlight alternate disease progression patterns in CKD. Irregularity and sparsity in longitudinal EHR data creates high variance in trajectory estimates and often leads to unstable clusters. The FCM approach addresses this challenge by treating curve coefficients as random effects, and then projecting the curve into a subspace where the cluster centers now represent the probability of cluster membership. Using this approach, the authors cluster creatinine trajectories of CKD patients to create two patient groupings which feature distinct clinical attributes.

2.3 Public Data Mining

The extraction and identification of higher level relationships from high-throughput data and data repositories is an important area of research. For example, with the ever increasing amount of study information existing within PubMed, it is a challenge to integrate that much information to gain higher level insights over trends that have been found for genes and diseases. The information gained from effectively integrating comprehensive data together in novel ways could ultimately result in the “sum being greater than the parts”, providing new insights for further research and discovery.

In “*A new relevance estimator for compilation and visualization of disease patterns and potential drug targets*”**, von Korff** et al. describe a tool, the Disease Relevance Miner (DDRelevanceMiner), which was developed using the concept of second order co-occurrence which takes advantage of calculating the similarity between two words that do not co-occur frequently, but co-occur with the same neighboring word. The authors used the basis of this approach but with the advancement of a relevance estimator. Using the DDRelevance Miner, the authors used HUGO gene identifiers, and then linked them to PubMed in order to extract relevant records for each gene, where each publication record in turn was searched with disease MeSH terms. Linking together these data along with a metric of relevance, provided detailed disease-gene and disease-disease associations which could be further explored. This includes the identification of gene drug targets that had indications of being highly specific to single diseases.

Wilson et al. evaluate the performance of four community detection algorithms to automatically determine groups of genes from protein-protein interaction networks using experimental data in "*Discovery of Functional and Disease Pathways by Community Detection in Protein-Protein Interaction Networks*". To date, biological pathway information has been based on experimentally gained understanding. The various pathway repositories that exist are incredibly important resources, a testament to how much has been learned of the underlying structure of biology. These resources contribute to a greater understanding of gene expression and genetic association results, as well as identification of genetic interaction candidates. High throughput computational approaches could help fast track the evaluation of new potential pathways. Determining communities of biological networks could shed new light on groupings of genes with common biological functions or features. With the reliance of many analyses based on gene and pathway information, such as the Gene Set Enrichment Analysis (GSEA) [11], Pathway Analysis by Randomization Incorporating Structure (PARIS) [12], and other tools like Biofilter [13], further identification of pathways could support new hypothesis generation for experimental validation. In the manuscript by Wilson et al., several possible community detection methods were tested using a STRING protein-protein interaction network [14]. Communities obtained were then compared to curated biological pathways, over multiple metrics. Both known pathways were re-identified and possibly novel pathways were identified, the authors carefully characterized other features of these networks as well, highlighting the utility of community detection methods in identifying new pathways for further study.

References:

1. Bourne PE, Bonazzi V, Dunn M, Green ED, Guyer M, Komatsoulis G, et al. The NIH Big Data to Knowledge (BD2K) initiative. J. Am. Med. Inform. Assoc. 2015;22:1114–1114.

2. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLoS Biol. 2015;13:e1002195.

3. Hvistendahl M. Can "predictive policing" prevent crime before it happens? Sci. Mag. [Internet]. 2016; Available from: http://www.sciencemag.org/news/2016/09/can-predictive-policing-prevent-crime-it-happens

4. Zhang Y, Zhao Y. Astronomy in the Big Data Era. Data Sci. J. 2015;14:11.

5. PMI in the News [Internet]. Natl. Inst. Health NIH. 2015 [cited 2016 Sep 23]. Available from: https://www.nih.gov/node/19706/draft

6. Rabes T, 1 ratanaAug, 2014, Pm 12:30. U.K.'s 100,000 Genomes Project gets £300 million to finish the job by 2017 [Internet]. Sci. AAAS. 2014 [cited 2016 Sep 23]. Available from: http://www.sciencemag.org/news/2014/08/uks-100000-genomes-project-gets-300-million-finish-job-2017

7. Carey DJ, Fetterolf SN, Davis FD, Faucett WA, Kirchner HL, Mirshahi U, et al. The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. Genet. Med. 2016;18:906–13.

8. Kaplan RM, Chambers DA, Glasgow RE. Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias. Clin. Transl. Sci. 2014;7:342–6.

9. DeRouen TA. Promises and Pitfalls in the Use of "Big Data" for Clinical Research. J. Dent. Res. 2015;94:107S – 109S.

10. Vadrevu S. Understanding the Promise and Pitfalls of Machine Learning [Internet]. Data Inf. 2015 [cited 2016 Sep 30]. Available from: http://data-informed.com/understanding-promise-pitfalls-machine-learning/

11. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. 2005;102:15545–50.

12. Butkiewicz M, Cooke Bailey JN, Frase A, Dudek S, Yaspan BL, Ritchie MD, et al. Pathway analysis by randomization incorporating structure-PARIS: an update. Bioinforma. Oxf. Engl. 2016;32:2361–3.

13. Pendergrass SA, Frase A, Wallace J, Wolfe D, Katiyar N, Moore C, et al. Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development. BioData Min. 2013;6:25.

14. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013;41:D808–15.

LEARNING ATTRIBUTES OF DISEASE PROGRESSION FROM TRAJECTORIES OF SPARSE LAB VALUES

VIBHU AGARWAL

Biomedical Informatics Training Program, Stanford University
Stanford, CA 94305, USA
Email: vibhua@stanford.com

NIGAM H SHAH

Center for Biomedical Informatics Research, Shah Lab, Stanford University
Stanford, CA 94305, USA
Email: nigam@stanford.com

There is heterogeneity in the manifestation of diseases, therefore it is essential to understand the patterns of progression of a disease in a given population for disease management as well as for clinical research. Disease status is often summarized by repeated recordings of one or more physiological measures. As a result, historical values of these physiological measures for a population sample can be used to characterize disease progression patterns. We use a method for clustering sparse functional data for identifying sub-groups within a cohort of patients with chronic kidney disease (CKD), based on the trajectories of their Creatinine measurements. We demonstrate through a proof-of-principle study how the two sub-groups that display distinct patterns of disease progression may be compared on clinical attributes that correspond to the maximum difference in progression patterns. The key attributes that distinguish the two sub-groups appear to have support in published literature clinical practice related to CKD.

1. Introduction

It is common knowledge that diseases manifest differently in different people. Knowing the alternative progression patterns of a disease for a given population, as well as the clinical attributes associated with the patterns, is therefore of interest to patients, doctors as well as researchers[1]. Knowing what to expect, empowers patients to make informed choices about their treatment options as well as plan a judicious acquisition of healthcare resources in the future. Furthermore, the ability to spot the unusual, and initiate a clinical evaluation in case the observed symptoms are anomalous with respect to known progression attributes, has the potential to improve the care delivery process. From the perspective of the care provider, knowing the attributes of the different paths of disease progression is essential for investigating risk factors associated with progression[2].

For a healthcare system preparing to care for an aging population, an understanding of disease paths as the basis for planning treatment can have a profound impact on the patient's wellness goals. For instance, it has been seen that classification of end-stage functional decline into four groups explains the observed patterns in a sample of older medicare decedents[3]. Insight into the most likely course of progression and the "signature" attributes, can prove invaluable to healthcare professionals. Finally, a knowledge of progression patterns is essential for discovering treatment options that alter disease progression. For instance, the stage duration as well as progression rates between normal aging and severe dementia, assessed via the Global Deterioration Scale in patients with Alzheimer's, show high heterogeneity[4]. A clinical evaluation of prospective therapies that seek to slow the cognitive decline in patients with Alzheimer's would need to be carried out in individuals with similar progression trends.

The general problem is of discovering patterns of clinical events associated with stages of progression and then classes of such sequential patterns. Generally, disease progression modelling efforts first learn a state transition model using comorbidity patterns and later infer the comorbidities that drive progression based on the observed symptoms[5]. However, for many diseases, the disease status can be reliably summarized by recording one or more physiological measures. Univariate measures such as Glycosylated Hemoglobin (Diabetes), Predicted Forced Vital Capacity (Scleroderma) and Estimated Glomerular Filtration Rate (Kidney Disease) are used routinely in medical practice. These measurements are typically recorded irregularly, and usually after long intervals, making the recorded trajectories sparse. For example, out of 18,342 patients with Type 2 Diabetes in our extract of patient data from the Stanford Clinical Data warehouse, only 8231 patients had two or more HbA1c measurements. The mean number of observations per patient was 7.49. An estimate of the disease progression path based on an observed trajectory of such measurements will have high variance. As a consequence, clusters derived from such path estimates are likely to be unstable.

We hypothesize that it is possible to learn clinically meaningful clusters of disease paths from sparse and irregular trajectories of lab values. In our earlier work[2] we have described a generative model for simultaneously modelling stages of progression in a cohort of Chronic Kidney Disease (CKD) patients, as well as discovering clusters of distinct progression sequence. In related work, there are prior efforts in creating finite dimensional representation of a trajectory captured by dense measurements, and cluster the trajectories using an appropriate similarity metric. Example of successful path estimation with methods employing Gaussian Process regression often involve measurements in post-operative care or the intensive care unit, where physiological measurements are recorded regularly and relatively few observations are missing[6–8].

In order to meaningfully cluster paths estimated from sparse measurement trajectories, it is possible to borrow support from other trajectories provided a large number of trajectories have been recorded for the full time grid. The Functional Clustering Model (FCM) proposed by James and Sugar[9] models sparse trajectories as random effects, after fitting natural cubic splines to observations from each trajectory. In the work presented here, we cluster creatinine measurements from patients with Chronic Kidney Disease using the FCM. We then compare the distribution of clinical features between patients in different clusters, by defining a time window around the region of maximum discrimination between the clusters. Finally, we examine the features whose distribution is significantly different between clusters, and interpret the differences in the light of published literature on the management of Chronic Kidney Disease. Figure 1 illustrates our approach and overall workflow.

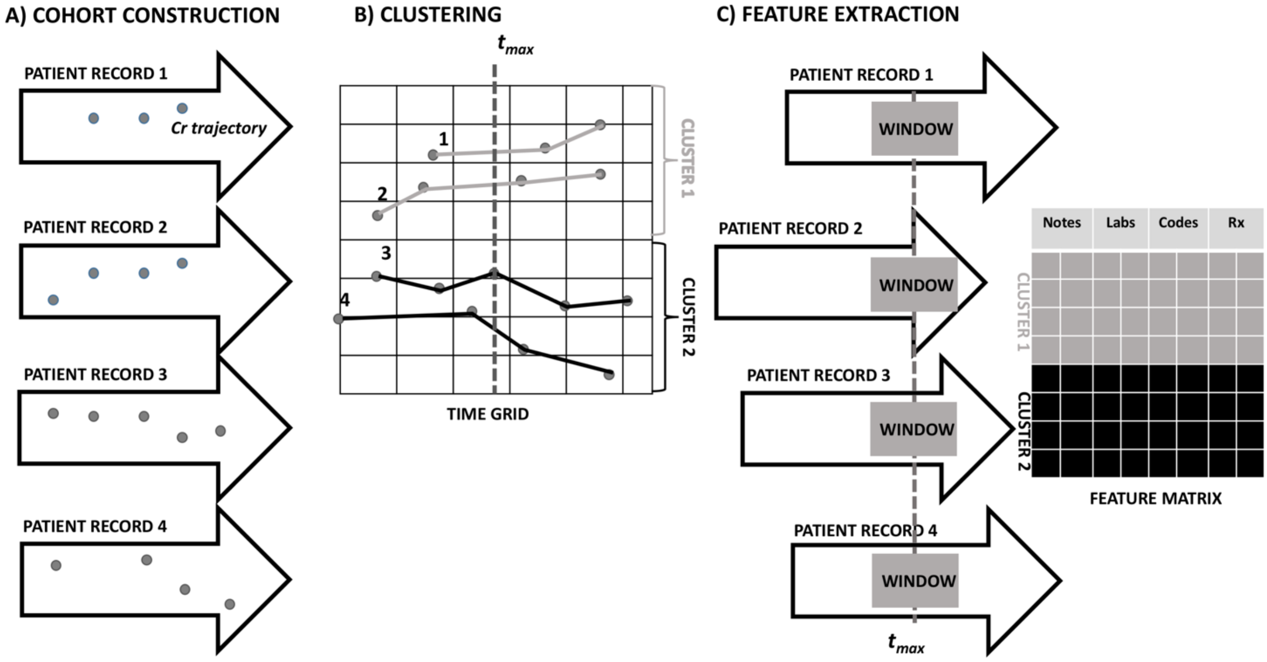


Figure 1A) Records for patients in the CKD cohort B) Clustering sparse trajectories of creatinine values. The time point at which the clustered trajectories are most discriminable is t_{max} C) Text concepts, lab results, ICD9 codes and prescriptions from a window centered at t_{max} in the patient records

2. Data

The patient dataset was extracted from the Stanford clinical data warehouse (SCDW), which integrates data from Stanford Children's Health (SCH) and Stanford Health Care (SHC). The extract comprises 2 million patients, with 49 million encounters, 35 million coded diagnoses and procedures, 204.8 million laboratory tests, 14 million medication orders as well as pathology, radiology, and transcription reports totaling over 27 million clinical notes. Our extract of the de-identified patient data from 01/1994 through 06/2013 from SCH and SHC is stored in a structured and indexed form within a MySQL relational database.

2.1. *Cohort selection*

In order to select patients with a diagnosis of Chronic Kidney Disease (CKD), we used the presence of the ICD-9 code 585.00 as our filtering criterion. We identified 959 CKD patients through this method, and kept 792 that had 3 or more creatinine measurements. We henceforth refer to this set of 792 patients as our cohort. Examining the sequence of creatinine measurements over time or "trajectories" from our cohort revealed significant sparsity in the creatinine measurements. Figure 2 shows the distribution of of the number of per patient measurements in our cohort, with a mean of 25.98 and median 13.

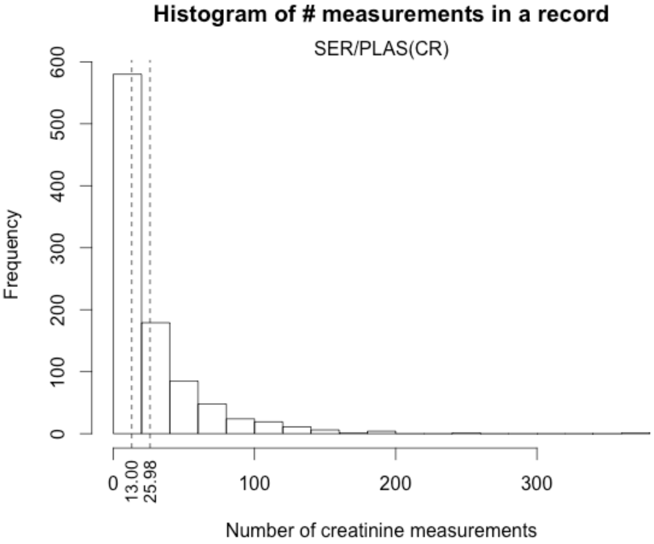


Figure 2 Distribution of number of measurements of creatinine per patient with the mean and median.

3. Methods

3.1. *Functional Clustering Model*

In order to cluster sparse observations in patient histories, we make use of the Functional Clustering Model (FCM)[9] which solves the problems of high variance due to sparsity as well as that of unequal variances due to irregular time instants. While a full description of the functional clustering approach and model fitting procedure is described by James and Sugar, the intuition behind the procedure is to project each curve onto a finite dimensional space using a natural cubic spline basis and cluster the resulting coefficients. However, instead of treating the basis coefficients as parameters James and Sugar model these are random effects that are ascertained by the clustering algorithm, via a global optimization over all curves. Doing so allows us to borrow strength across curves, allowing the method to work with sparsely or irregularly sampled curves, provided that the total number of observations is large enough.The model is fit to the data using an Expectation Maximization like procedure that iteratively updates the FCM parameters. As shown in Figure 3, each track is represented as a vector of measurements $\boldsymbol{Y_i}$ such that

$$\boldsymbol{Y_i} = \boldsymbol{g_i} + \boldsymbol{\epsilon_i} \tag{1}$$

where g_i represents the true (unobserved) measurements at the same instants and ϵ_i is the vector of random observation errors $\epsilon_i \sim N(0,\ \sigma^2 I)$. The observation errors are assumed to be uncorrelated with each other and with g_i. We assume membership in one among G clusters. The true observations are modelled by using natural cubic splines as basis functions $s(t)$ so that

$$g_i(t) = s(t)^T \eta_i \tag{2}$$

where η_i is the vector of spline coefficients. The FCM uses a random effects model for the η_i, assuming a normal distribution of values around a cluster mean as follows

$$\eta_i = \mu_{z_i} + \gamma_i \tag{3}$$

where μ_{z_i} represents the mean of the i^{th} cluster of tracks and $\gamma \sim N(0,\ \Gamma)$. An additional parameterization step allows for a low dimensional representation of η_i given by

$$\eta_i = \lambda_0 + \Lambda\alpha_{z_i} + \gamma_i \tag{4a}$$
$$\text{where } \mu_k = \lambda_0 + \Lambda\alpha_k \tag{4b}$$

Thereafter the FCM can be represented as below

$$Y_i = Si^T(\lambda_0 + \Lambda\alpha_{z_i} + \gamma_i) + \epsilon_i \tag{5}$$

where Si is the matrix of basis expansions of all time points for track i, α_{z_i} is a h vector representing the mean of the i^{th} cluster of trajectories in a low dimensional space. The low dimensional representation is achieved via $\mu_k = \lambda_0 + \Lambda\alpha_k$ where both μ_k, λ_0 are vectors in $\mathbb{R}^p$, Λ is a p x h matrix and $h \leq min(p, G\text{-}1)$. To ensure a unique solution for μ_k, λ_0 and Λ, we require

$$\sum \alpha_i = 0 \tag{6a}$$
$$\Lambda^T S^T (\sigma^2 I + S\Gamma S^T)^{-1} S\Lambda = I \tag{6b}$$

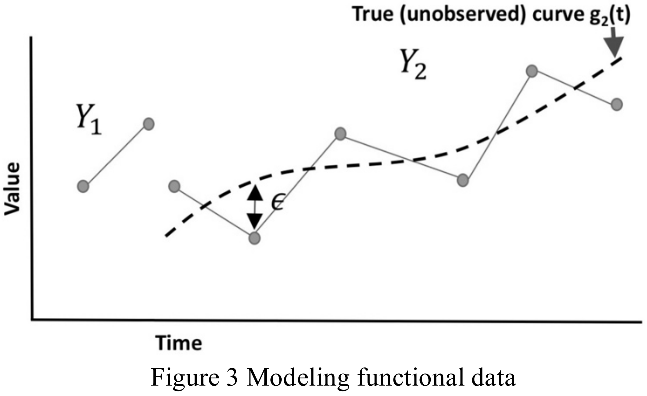


Figure 3 Modeling functional data

The form of (6b) ensures that $Cov(\alpha_i) = I$ for all i, when the observations for each track are measured at the same time points. The above formulation allows us to project every trajectory into h dimensional space to obtain the corresponding $\hat{\alpha}_i$, such that the proximity between $\hat{\alpha}_i$ and α_k represents the likelihood that the track i belongs to cluster k. Since $\hat{\alpha} = \Lambda^T S^T (\sigma^2 I + S\Gamma S^T)^{-1}(Y - S\lambda_0)$, the vector $\Lambda^T S^T (\sigma^2 I + S\Gamma S^T)^{-1}$ may be thought of as weights that determine the proximity to α_k, with the highest weight having the most influence on cluster assignment. This allows us to use $\Lambda^T S^T (\sigma^2 I + S\Gamma S^T)^{-1}$ as a discriminant function, for determining the time points that provide maximum cluster discrimination.

We provide here only an overview of the FCM and point the reader to a full description of the model, constraints and the fitting algorithm[9]. Since intuitively one expects to see a small number of distinct trajectory patterns, we clustered the creatinine measurements using small values of G. We obtained two nearly indistinguishable clusters with G=3, which suggests that G=2 may be an appropriate number of clusters for the creatinine trajectories in our cohort. After fitting the FCM with G=2 and making cluster assignments for every trajectory, we plotted the discriminant function for the two clusters in order to distinguish the time at which the two groups of trajectories most differ from each other.

3.2. *Feature Engineering and Analysis*

For each patient record in our cohort, we construct a one-year time window centered around the maximum discriminating time point as identified by the discriminant function described earlier. Within the time window, we represent the structured and unstructured data within a patient record as features from four categories – terms (or concepts), prescriptions, laboratory test results and diagnosis codes. Prescriptions, laboratory test results and diagnosis codes were taken from the structured record whereas terms were extracted from free text. We normalize terms into concepts in the same manner as in our earlier studies involving text mining on clinical notes—essentially using UMLS term-to-concept maps with suppression rules to weed out ambiguous mappings as described by Jung et al[10] . Such mapping reduces the total number of features as well as reduces the number of correlated features since synonyms get mapped to the same concept.

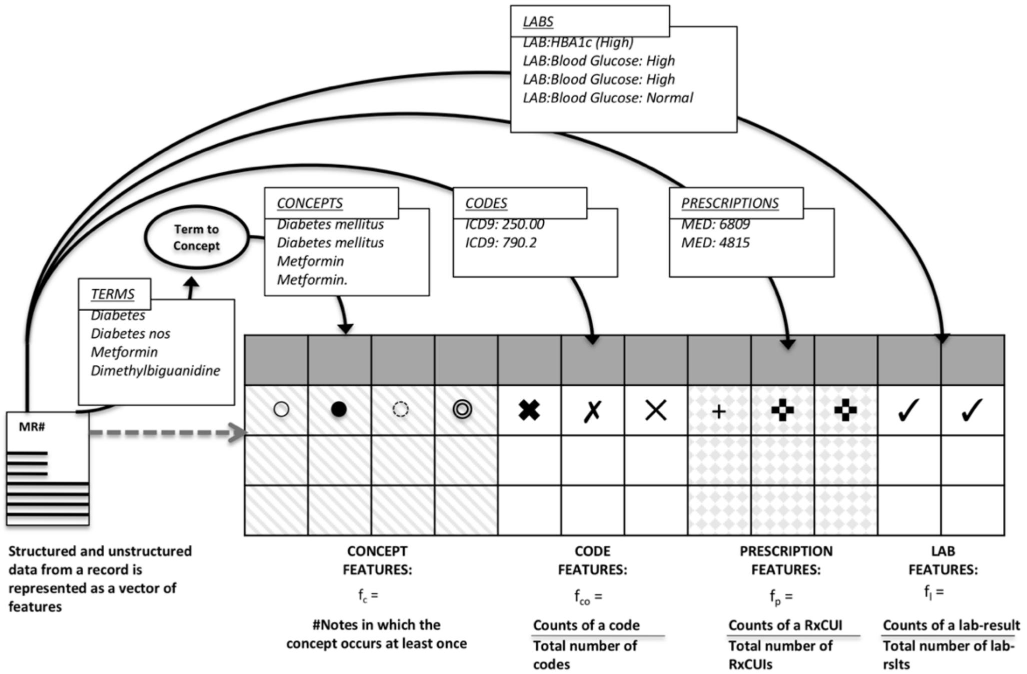


Figure 4 Features based on normalized counts of text concepts, ICD9 codes, prescriptions and lab results from the 1 year window around the maximum discriminating time point in the patient record

For concepts, we used the number of distinct notes in which the concept occurs at least once (note frequency) as the feature representation. For prescriptions and diagnostic codes, we used the normalized counts of the active ingredient for each medication (RxNorm concept unique identifier) and the normalized counts of each International Classification of Diseases, revision 9 (ICD9) code as the respective features. For laboratory test results, we utilized the categorical result status for each ordered test (high/ normal/low or normal/abnormal) as recorded in the Electronic Health Record (EHR) and calculated a feature based on the normalized counts for each test-result instance in the record. Our feature construction method is illustrated in Figure 4 which depicts our feature matrix along with the four categories (concepts, diagnosis codes, prescriptions and laboratory results) of data elements within the patient record from which the features are sourced. Finally, we performed an enrichment analysis on the feature matrix for the two clusters using Fisher's exact test, setting a false discovery rate of 5% to adjust for multiple testing. All analysis was performed in R using the Aphrodite[11] and the fclust[12] APIs.

4. Results

Figure 5 shows a randomly selected subset of 50 trajectories from each of the two clusters, indicating the overall progression pattern in each cluster. The thick lines corresponding to each cluster mean show progression patterns that the respective cluster represents. In case of cluster 1, the mean trajectory indicates that creatinine levels begin to rise around the age of 65 years, peak at around 72 years and then decrease. The trajectories in cluster 2 suggest an overall better control on creatinine levels.

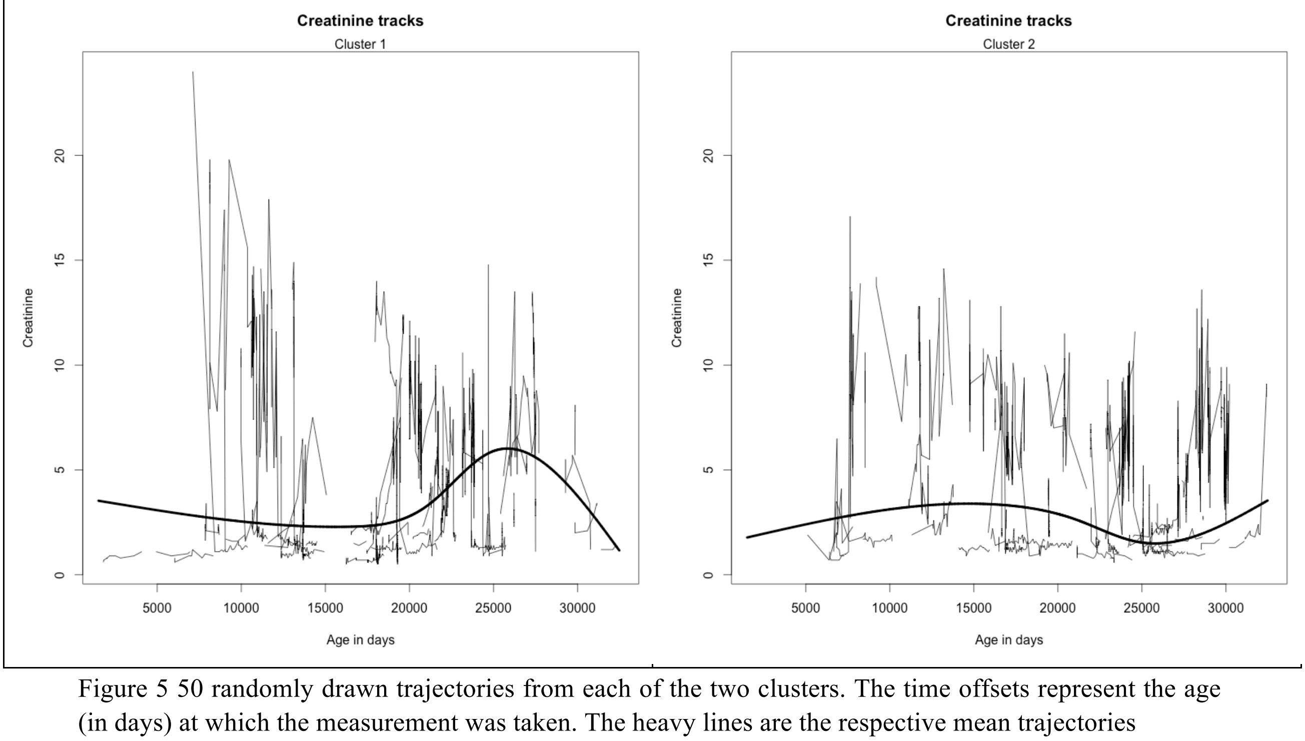


Figure 5 50 randomly drawn trajectories from each of the two clusters. The time offsets represent the age (in days) at which the measurement was taken. The heavy lines are the respective mean trajectories

The maximum value of the discriminant function occurs at a time offset t_{dmax} = 26,383 days (72 years), corresponding to the peak in the mean Creatinine trajectory in cluster 1. Drawing concepts, ICD9 codes, prescriptions and lab results from the EHR records of patients in each of the two clusters for the 1-year time window centered on t_{dmax} yields 1046 features. The rate of progression to end stage renal disease depends on a number of risk factors, such as the presence of CKD-associated illnesses, nutrition issues and adherence to medication[13] and we expect features from 1 year window to represent events associated with clinically significant decline in renal function in advanced-stage patients[14].

Table 1 Top ranked significant features

Rank	Feature ID (p)	Names	Rank	Feature ID (p)	Names
1	lab.3023314.0 (3.27e-05)	Hematocrit [Volume Fraction] of Blood by Automated count:	16	obs.4274025 (7.86e-04)	Disease
2	lab.3000963.0 (4.06e-05)	Hemoglobin:	17	obs.4322976 (8.01e-04)	Procedure
3	lab.3000905.0 (5.92e-05)	Leukocytes [#/volume] in Blood by Automated count:	18	obs.4229881 (8.46e-04)	Weight loss
4	obs.4187395 (7.19e-05)	Reflux	19	obs.46233416 (9.53e-04)	Assessment
5	obs.4187458 (1.46e-04)	Review of systems	20	obs.4143467 (1.09e-03)	Chief complaint
6	obs.4243768 (1.78e-04)	Auscultation	21	lab.3009261.0 (1.22e-03)	Glucose [Presence] in Urine by Test strip:
7	obs.4118663 (2.06e-04)	Related	22	obs.4209224 (1.51e-03)	Cyst
8	obs.31967 (3.23e-04)	Nausea	23	obs.441408 (1.67e-03)	Vomiting
9	lab.3013682.0 (3.25e-04)	Urea nitrogen serum/plasma:	24	obs.4147571 (1.82e-03)	Follow-up
10	obs.442985 (3.85e-04)	Male	25	obs.4267147 (1.85e-03)	Platelet count
11	lab.3014051.0 (4.03e-04)	Protein [Presence] in Urine by Test strip:	26	lab.3022621.0 (1.85e-03)	pH of Urine by Test strip:
12	obs.254761 (4.21e-04)	Cough	27	obs.77670 (1.86e-03)	Chest pain
13	obs.4077953 (4.88e-04)	Therapy	28	lab.3035350.0 (1.86e-03)	Ketones urine dipstick:
14	obs.4099313 (5.19e-04)	Urinalysis	29	lab.3004501.45876384 (1.96e-03)	Glucose lab:High
15	obs.4329041 (7.37e-04)	Pain	30	obs.4303558 (1.98e-03)	Touch

A Fisher's exact test of enrichment for each feature with respect to the two clusters, using a false discovery rate of 5% to adjust for multiple testing, identified 133 enriched features, of which 30 top ranked features along with their respective p-values are presented in Table 1.

All of the top 30 features are found to be enriched in cluster 1 and an examination of the features suggests concordance with the known attributes of disease severity. For example the top 2 features (lab orders for hematocrit and hemoglobin) suggest the presence of Anemia and Thrombocytopenia respectively, which are two common comorbidities in advanced CKD that are thought to occur as a result of reduced erythropoietin secretion[15]. An observation related to platelet counts (feature rank 25) corroborates this view. Similarly, automated blood count is routinely measured in CKD patients, particularly in those patients requiring management of Anemia. Recently studies show that spikes in granulocyte and monocyte count in CKD patients are associated with progression to end stage renal disease[16]. A finding of Proteinuria on dipstick urinalysis is amongst the early signs of kidney disease, however high levels of protein in the urine is an indicator of nephritic syndrome and is associated with edema, increased cholesterol levels and other comorbidities that increase the risk of CKD progression[17]. Lab orders for pH and ketone measurement in urine (feature rank 26 and 28 respectively) suggest diabetic ketoacidosis which is an uncommon but life threatening complication in chronic kidney disease. Poor glucose regulation (feature rank 29) further supports the view that several of the distinguishing attributes have etiological linkages with diabetic complications.

5. Discussion

Longitudinal patient data from a large sample of patients offers an opportunity to characterize the variability in how phenotypes progress over time. However, discovering patterns of progression for chronic diseases is challenging because of the irregularity and sparsity in the observations. Trajectory estimates derived from irregularly sampled and sparse observations have high variance which leads to unstable cluster definitions. The irregular sampling also poses an additional challenge – the variance of the estimated curve coefficients are different for each trajectory. The FCM described in the methods section addresses the problem by treating the curve coefficients as random effects and by projecting each curve into a subspace, such that the covariance normalized distance from the cluster center in this subspace, represents the probability of cluster membership. Using the FCM to cluster creatinine trajectories of CKD patients results in two clusters with distinct mean trajectories. Features based on counts of clinical attributes, from a windowed segment of the patients' EHR records at the point of maximal cluster separation, show a significantly different distribution in the two clusters; and many are supported by medical literature on CKD. However, several of our significant features refer to general CKD attributes and do not appear to have an obvious connection with disease severity or progression. Further, reducing the false discovery rate to 1% did not give any statistically significant features.

We acknowledge limitations of our approach. We made use of the ICD9 code for CKD for defining our cohort. Such a method could have a positive predictive value (PPV) as low as 53%[18] if relying solely on the ICD9 code. In which case, the the selected patients may not have the clinical indicators for the disease, while creatinine measurements could still be available since creatinine may be ordered routinely as part of a basic metabolic panel. We mitigate this issue by requiring at least three creatinine measurements for members of the analysis cohort. Further still, creatinine values are known to be altered through processes that are independent of renal function and standard practice requires that the estimated Glomerular Filtration Rate be used for assessing kidney health[19]. The FCM also implicitly assumes that the unobserved time points are missing at random. Given that our data comes from a referral facility, it is likely that there exists a "disease severity bias" in the missing observations.

The alternative to using ICD9 codes for cohort identification is to use a robust algorithm that has been validated to achieve a high PPV[20]. Patient records extracted from claims data may possibly provide better longitudinal coverage compared to EHR data from a tertiary care facility. Addressing our study's limitations through the aforementioned remedial measures appears feasible and we anticipate doing so in follow up work.

6. Conclusion

Being able to account for individual variability in the progression of diseases is of value to the practitioner, the patient as well as the researcher. For chronic diseases, learning the clinical attributes of the disease progression paths is possible by using a method for clustering irregularly sampled, sparse trajectories of disease markers, by defining a time window in which the clusters are most discriminable, and identifying discriminating features based on that time window. Our results from clustering creatinine trajectories of CKD patients demonstrate the feasibility of the approach.

7. Acknowledgements

We would like to thank our colleagues Sarah Poole and Jassi Pannu for their contributions to this study. This work was funded by NLM R01 LM011369.

References

1. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet*. 2012;13(6):395-405. doi:10.1038/nrg3208.
2. Yang J, McAuley J, Leskovec J, LePendu P, Shah N. Finding progression stages in time-evolving event sequences. *Proc 23rd Int Conf World wide web*. 2014:783-794. doi:10.1145/2566486.2568044.
3. Lunney JR, Lynn J, Hogan C. Profiles of Older Medicare Decedents. *J Am Geriatr Soc*. 2002;50(6):1108-1112. doi:10.1046/j.1532-5415.2002.50268.x.
4. Komarova NL, Thalhauser CJ. High degree of heterogeneity in Alzheimer's disease progression patterns. *PLoS Comput Biol*. 2011;7(11):e1002251. doi:10.1371/journal.pcbi.1002251.
5. Wang X, Wang F. Unsupervised Learning of Disease Progression Models. 2014. doi:10.1145/2623330.2623754.
6. Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLoS One*. 2013;8(6):e66341. doi:10.1371/journal.pone.0066341.
7. Pimentel M, Clifton D, Clifton L, Tarassenko L. Modelling Patient Time-Series Data from Electronic Health Records using Gaussian Processes. *Adv neural Inf Process Syst Work Mach Learn Clin Data Anal*. 2013:1-4.
8. Pimentel, MAF, Clifton DA TL. Gaussian process clustering for the functional characterisation of vital-sign trajectories. In: *Machine Learning for Signal Processing*

(MLSP), 2013 IEEE International Workshop On. ; 2013:1-6. doi:10.1109/MLSP.2013.6661947.

9. James GM, Sugar C a. Clustering for Sparsely Sampled Functional Data. *J Am Stat Assoc*. 2003;98(462):397-408. doi:10.1198/016214503000189.
10. Jung K, LePendu P, Iyer S, Bauer-Mehren A, Percha B, Shah NH. Functional evaluation of out-of-the-box text-mining tools for data-mining tasks. *J Am Med Inform Assoc*. 2015;22(1):121-131. doi:10.1136/amiajnl-2014-002902.
11. OHDSI. Aphrodite. https://github.com/OHDSI/Aphrodite. Accessed October 3, 2016.
12. Gareth J. Fclust. http://www-bcf.usc.edu/~gareth/research/fclustdoc.pdf. Accessed October 3, 2016.
13. Thomas R, Kanso A, Sedor JR. Chronic kidney disease and its complications. *Prim Care*. 2008;35(2):329-344, vii. doi:10.1016/j.pop.2008.01.008.
14. Zhang A-H, Tam P, LeBlanc D, et al. Natural history of CKD stage 4 and 5 patients following referral to renal management clinic. *Int Urol Nephrol*. 2009;41(4):977-982. doi:10.1007/s11255-009-9604-3.
15. Akimoto T, Ito C, Kotoda A, et al. Challenges of caring for an advanced chronic kidney disease patient with severe thrombocytopenia. *Clin Med Insights Case Rep*. 2013;6:171-175. doi:10.4137/CCRep.S13238.
16. Agarwal R, Light RP. Patterns and prognostic value of total and differential leukocyte count in chronic kidney disease. *Clin J Am Soc Nephrol*. 2011;6(6):1393-1399. doi:10.2215/CJN.10521110.
17. Mehdi U, Toto RD. Anemia, diabetes, and chronic kidney disease. *Diabetes Care*. 2009;32(7):1320-1326. doi:10.2337/dc08-0779.
18. Cipparone CW, Withiam-Leitch M, Kimminau KS, Fox CH, Singh R, Kahn L. Inaccuracy of ICD-9 Codes for Chronic Kidney Disease: A Study from Two Practice-based Research Networks (PBRNs). *J Am Board Fam Med*. 28(5):678-682. doi:10.3122/jabfm.2015.05.140136.
19. Samra M, Abcar AC. False estimates of elevated creatinine. *Perm J*. 2012;16(2):51-52.
20. Nadkarni GN, Gottesman O, Linneman JG, et al. Development and validation of an electronic phenotyping algorithm for chronic kidney disease. *AMIA Annu Symp Proc*. 2014;2014:907-916.

COMPUTER AIDED IMAGE SEGMENTATION AND CLASSIFICATION FOR VIABLE AND NON-VIABLE TUMOR IDENTIFICATION IN OSTEOSARCOMA

HARISH BABU ARUNACHALAM[1]*, RASHIKA MISHRA[1], BOGDAN ARMASELU[1],
DR. OVIDIU DAESCU[1] and MARIA MARTINEZ[2]

[1] *Department of Computer Science,*
[2] *Department of Biomedical Engineering,*
University of Texas at Dallas, Richardson, TX
**Email: harishb@utdallas.edu*

DR. PATRICK LEAVEY, DR. DINESH RAKHEJA, DR. KEVIN CEDERBERG,
DR. ANITA SENGUPTA and MOLLY NI'SUILLEABHAIN
University of Texas Southwestern Medical Center, Dallas, TX

ABSTRACT: Osteosarcoma is one of the most common types of bone cancer in children. To gauge the extent of cancer treatment response in the patient after surgical resection, the H&E stained image slides are manually evaluated by pathologists to estimate the percentage of necrosis, a time consuming process prone to observer bias and inaccuracy. Digital image analysis is a potential method to automate this process, thus saving time and providing a more accurate evaluation. The slides are scanned in Aperio Scanscope, converted to digital Whole Slide Images (WSIs) and stored in SVS format. These are high resolution images, of the order of 10^9 pixels, allowing up to 40X magnification factor. This paper proposes an image segmentation and analysis technique for segmenting tumor and non-tumor regions in histopathological WSIs of osteosarcoma datasets. Our approach is a combination of pixel-based and object-based methods which utilize tumor properties such as nuclei cluster, density, and circularity to classify tumor regions as viable and non-viable. A K-Means clustering technique is used for tumor isolation using color normalization, followed by multi-threshold Otsu segmentation technique to further classify tumor region as viable and non-viable. Then a Flood-fill algorithm is applied to cluster similar pixels into cellular objects and compute cluster data for further analysis of regions under study. To the best of our knowledge this is the first comprehensive solution that is able to produce such a classification for Osteosarcoma cancer. The results are very conclusive in identifying viable and non-viable tumor regions. In our experiments, the accuracy of the discussed approach is 100% in viable tumor and coagulative necrosis identification while it is around 90% for fibrosis and acellular/hypocellular tumor osteoid, for all the sampled datasets used. We expect the developed software to lead to a significant increase in accuracy and decrease in inter-observer variability in assessment of necrosis by the pathologists and a reduction in the time spent by the pathologists in such assessments.

Keywords: Image segmentation, Otsu thresholding, Osteosarcoma, SVS image analysis

1. Introduction

Pathology Informatics, one of the fastest growing fields in medical informatics, deals with mining information from medical pathology data and images. It involves the use of computational methods and analytical processes to make informed decisions that serve as assistive tools in clinical diagnosis. Due to the complexity of medical data and given the expert knowledge required for such analyses, it is often difficult to replicate the work of pathologists and physicians.[1] Though there is substantial literature published in the area of tumor research[2,3] the main challenge in the field is that all the methods are tumor specific which makes the development of one common method, that is applicable for all kinds of tumor, an arduous task. This necessitates the creation of ad hoc methods tied to each requirement, that consider signature of each tumor sample and incorporate tumor specific information such as the tumor spread, contextual information etc. Each tumor detection method utilizes specific information about the tumor and therefore one tumor identification approach may not be applicable for another. Hence it becomes a challenge to apply existing methods that work well for other types of tumors for Osteosarcoma detection, the tumor that is used in this study.

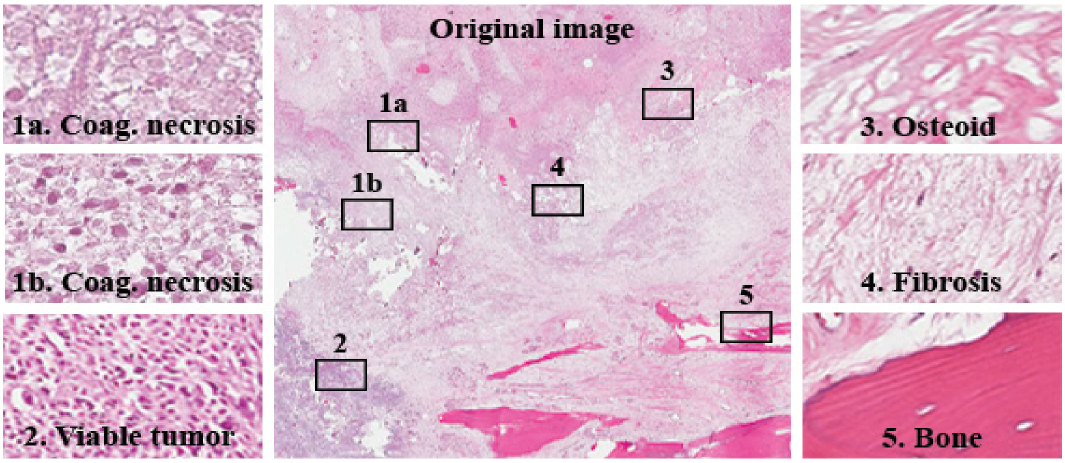


Fig. 1. WSI with different regions enlarged and their location in the image with figure 1a and 1b representing color and shape variations in coagulative necrosis regions for the same WSI. The numbered boxes represent the locations of histologically distinct regions in the image.

Osteosarcoma is the most common type of bone cancer that occurs in adolescents in the age of 10 to 14 years. The tumor usually arises in the long bones of the extremities in the metaphyses next to the growth plates.[4] What makes osteosarcoma analysis inherently challenging is that there is a high degree of histologic variability within the tumor (Figure 1) which is accentuated after therapy. In order to accurately identify tumor occurrence and estimate the treatment response, it is necessary to consider histologically distinct regions that include dense clusters of nuclei, fibrous tissues, blood cells, calcified bone segments, marrow cells, adipocytes, osteoblasts, osteoclasts, haemorrhagic tumor, cartilage, precursors, growth plates and osteoid (tumor osteoid and reactive osteoid) with and without cellular material. Each of these regions have different characteristic features that differ in color, shape, size, density, texture and area of occurrence. They also have significant differences in their biological features such as background stroma, presence or absence of certain cellular material, neighboring regions etc. There are also multiple color variations within the same dataset representing the same type of regions (Figure 1a and 1b), which makes segmentation based only on color ineffective. Due to the variable properties of the images, there is no one method that with certainty, can accurately segment regions and classify them. The literature available for osteosarcoma data image analysis/digital pathology is minimal which makes it critical to come up with methods that are applicable for this type of tumor analysis.

The goal of this paper is to present an approach to segment H&E[5] stained images into viable tumor and non-viable tumor regions using a combination of techniques (color segmentation, Otsu thresholding, and Flood-fill). It employs pixel-based and object-based segmentation by using color, shape and density parameters. The first step locates tumor and non-tumor regions and subsequent steps distinguish tumor into viable and non-viable regions. Color quantization in the approach accounts for variance in color distribution while shape properties such as area and circularity measures are utilized to accurately locate tumor. A further analysis performed on computed cluster data on resulting images characterize different regions in the image. The approach is shown to be robust and has a high accuracy overall in the datasets considered.

1.1. *Background and Setup*

We have assembled an investigative team of clinical scientists at University of Texas Southwestern Medical Center, Dallas and computer scientists at University of Texas at Dallas. Archival samples for 50 patients treated at Children's Medical Center, Dallas, between 1995 and 2015, have been identified. The treatment effect for each patient is estimated after surgical resection, by physically cutting the region of interest from the resected bone into pieces. These pieces are de-calcified, treated with H&E stain and converted to slides to be analyzed under microscope. Each patient case is represented by a single H&E stained slide at the time of biopsy when available and 8-50 H&E stained slides per case at time of resection when necrosis is determined. Each slide consists of 1-2cm X 1-2cm sections of the tumor in its widest coronal plane. Slides are scanned using an Aperio Scanscope at a magnification of upto 40X and stored in SVS format.[6] Each SVS WSI has a size between 150MB and 1.5GB and spans an order of 10^9 pixels. The experimental dataset consists of a subset of images from the above available patient datasets, manually annotated by pathologists.

1.2. *Related Work*

The Tumor identification/isolation problem has been well studied by researchers in the field of digital pathology[7–11] . The most common methods include image segmentation (region classification), image analysis (pattern analysis), regions of interest(ROI) identification, statistical analysis (such as number of clusters, mean size of groups) etc.[7,8] The methods used in all the studies include color based (pixel level), shape based (object level) and contextual information based methods.[8] Normally, pixel level methods, that make use of pixel level processing, form some of the basic approaches because they are the simplest but they are not the most efficient. Researchers have tried to analyze the pathological images based on quantitative metrics representing the spatial structure of histopathology imagery and include identifying structures such as nuclei, glands and lymphocytes etc. These spatial feature utilization[12] has become the backbone of histopathology image analysis techniques, as these are the prominent metrics that can yield maximum information. More advanced than the pixel based methods are the object based methods that make use of region growing and object identification utilizing shape properties. These methods provide better segmentation results than their pixel counterparts[2,13] however they are expensive in terms of the computational resources they use. Multi-level thresholding approach using Otsu segmentation promises good accuracy but when there is a lot of noise, this method alone may not fare well. Pattern Recognition Image Analysis(PRIA) describes a pattern recognition method based on a genetic algorithm that evolves over multiple iterations and compares the results with GeNIE,[14] a bio-image analysis tool from Aperio and manual segmentation by pathologists. A recent work on Lung cancer[11] identifies 9879 image features and uses regularized classifiers to estimate patient prognosis.

1.3. *Challenges*

A majority of pathological images are in proprietary format and there is no common standard for these images, which makes it difficult for researchers collaborating towards common goals to share information. Openslide[15] and VIPS[16] are image processing libraries that help to narrow down the gap by a small margin, however, the main problem of handling different imaging formats remains the same.

Pathology is a relatively subjective field dependent on the opinions of trained pathologists resulting in discrepancy in the accuracy of different image analysis approaches available for pathological images. A study on renal cell carcinoma[10] performed analysis on pathologists and found a high degree of subjectivity in their evaluation. Therefore, a standard objective procedure is recommended, which is important not only from a clinical standpoint but also for developing quality research application that will be reliable and independent of varying views of pathologists.[7] The size of images generated in these studies is large and as a result the algorithms for smaller images may not scale up due to memory issues.[17] Hence, there is a need for a standard approach that will process images one tile at a time and at the same time can scale up without loss of accuracy. Most of the tools that are developed by researchers in academia cater to a subset of problems. ImageJ[18] has many inbuilt image processing algorithms, however, is limited in its use to process proprietary formats and large files. PRIA by Webster et.al[1] is an advanced method, but it fails to perform well in identifying necrosis, which is one of the main tasks in this study. CellProfiler,[19] a tool for high throughput image analysis is good at identifying cellular objects and calculating their properties. However, the results of our trials with CellProfiler are inconclusive in identifying cellular objects in Osteosarcoma. Object based methods[2,13] work well on images with well-defined shapes but need pre-configured training sets. Machine learning approaches such as Bayesian classifier, Support Vector Machines[9] etc. are effective but would need a large annotated training data and the training phase is very time consuming.[7] Some of the related works[9–11] in lung and breast cancer focus on identifying properties and features of nuclei. Necrotic regions in this study do not necessarily have nuclei and hence the above works address only a part of the problem. The presence of a high degree of variability in the shapes, in Osteosarcoma datasets, makes the above methods unlikely to perform well. Another issue with WSI images is the color variance between different features,[9] which causes active tumor cells to have different color signatures, and thus segmentation based only on color is less accurate.

Given these challenges, in the next few sections we describe our approach explaining the algorithm and our results.

2. The Approach

We illustrate in Figure 2, the complete procedure, which includes K-means color segmentation, multi-level Otsu segmentation, Flood-fill clustering and statistical analysis. Based on inputs from pathologists, we define the different tumor regions with the following properties which are then quantified in the segmentation and classification approach.

(1) **Viable Tumor:** Nuclei densely aggregated together

(2) **Non-Viable Tumor:** Cells and tissues in the stage of recovery or dead

  (a) *Coagulative Necrosis:* disintegrated nuclei but with less color density than viable tumor.

  (b) *Fibrosis:* Fibrous collagen (protein) produced by fibroblasts (benign cells).

  (c) *Acellular/Hypocellular Tumor Osteoid (subsequently designated as Osteoid in this paper):* Eosinophilic/pink extracellular protein matrix produced by tumor cells.

The viable tumor and coagulative necrosis regions resemble each other in terms of high color density, closer to blue while fibrosis and osteoid have brighter color shade, closer to pink. The above regions are grouped together into two intermediary classes Ψ_1 and Ψ_2 based on high intra-class similarities, as follows:

(1) $\Psi_1 = \{Viable\ tumor, Coagulative\ necrosis\}$

(2) $\Psi_2 = \{Fibrosis,\ Osteoid\}$

The images in Ψ_1 are analyzed in terms of shape and density properties to classify them as viable tumor and non-viable tumor, while those in Ψ_2 are by default classified as non-viable tumor.

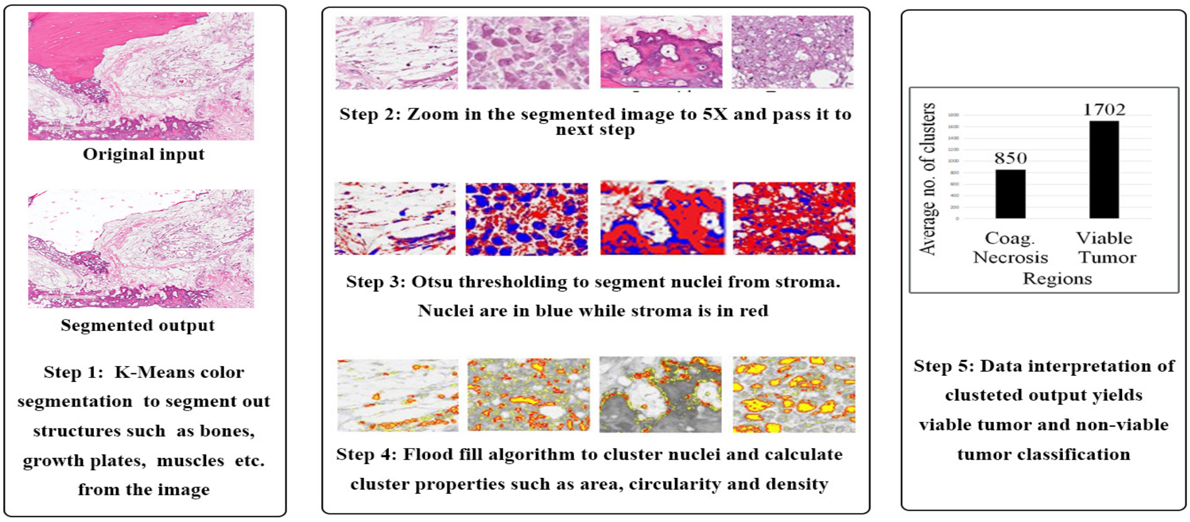

Original input
Segmented output
Step 1: K-Means color segmentation to segment out structures such as bones, growth plates, muscles etc. from the image
Step 2: Zoom in the segmented image to 5X and pass it to next step
Step 3: Otsu thresholding to segment nuclei from stroma. Nuclei are in blue while stroma is in red
Step 4: Flood fill algorithm to cluster nuclei and calculate cluster properties such as area, circularity and density
Average no. of clusters
1800
1600
1400
1200
1000
800
600
400
200
0
850
1702
Coag. Necrosis
Viable Tumor
Regions
Step 5: Data interpretation of clusteted output yields viable tumor and non-viable tumor classification

Fig. 2. **Algorithm pipeline**

2.1. *The Algorithm Pipeline*

The following is the general algorithm pipeline.

Input: Unprocessed SVS image

Output: Color segmented image and mapped regions identified as tumor (viable and non-viable) and non-tumor

Steps

For each SVS image given as input, at the eye fit (scaling factor 1X) zoom level, do the following:

1. Run K-means color segmentation with K=3.
2. For each of the tumor regions identified in step 1, increase the scaling factor to 5X.
3. On a window size of 512 * 512, on the original 5X scaled images, do the following:
    - (a) Generate red-blue segmentation using 2-level Otsu thresholding.
    - (b) Compute the percentage of red pixels and blue pixels in each image.
    - (c) Images with higher percentage of blue pixels fall into the Ψ_1 class
    - (d) Images with higher percentage of red pixels fall into the Ψ_2 class
    - (e) Create a tumor map of the pixel information by including pixel location, color values and class label
4. For each entry in the tumor map,
    - (a) Run Flood-fill algorithm to identify boundaries and group pixels in the cluster.
    - (b) Remove clusters that are smaller than minimum cluster size and larger than maximum cluster size to remove false positives.
    - (c) Output the remaining pixels in the original image along with cluster-mapped pixels.
5. Run data analysis and classify images as viable and non-viable tumor based on the cluster data output.

2.2. *Color Segmentation*

At eye-fit level(scaling factor 1X), a 3-means color segmentation process is used to distinguish the given image into tumor and non-tumor image. Since all the WSIs are H&E stained, the pixels are made up of variants of Red and Blue channels. Hence it is imperative that more focus is given to these two channels. Given an Image I, of width I_w and height I_h, made up of $I_w * I_h$ pixels, each pixel P^i is represented by $\{P^i_r, P^i_g, P^i_b\}$, where P^i_r, P^i_g and P^i_b denote the red, blue and green channel values of i^{th} pixel. Let the set $C \in (C^w, C^p, C^b)$ represent the cluster centers of white, pink and blue regions. These color values are taken from the empirical analysis of the stained images. Each P^i in the image is assigned to one of the cluster centers C^k by calculating distances between the pixel and the centroids. The distance is given by subtracting the color channel differences between the pixel and centroid. If $\phi(P^i)$ represents the cluster value for pixel i, then

$$\phi(P^i) \ = \underset{C^j}{\arg\min} \ \delta(P^i, C^j) \tag{1}$$

where,

$$\delta(P^i, C^j) \ = \ \sqrt{(P^i_r - P^k_r)^2 \ + \ (P^i_g - P^k_g)^2 \ + \ (P^i_b - P^k_b)^2} \tag{2}$$

where P^k_r, P^k_g, P^k_b represent RGB color channel values of the pixel, of k^{th} cluster centroid. The centroids are initialized with random values and each pixel P^i in I is classified. The pixel values of centroids are then updated as follows:

$$P_r^k = \frac{1}{N_k}\sum_j P_r^j;\ P_g^k = \frac{1}{N_k}\sum_j P_g^j;\ P_b^k = \frac{1}{N_k}\sum_j P_b^j; \tag{3}$$

Where N_k is the number of pixels classified under k^{th} centroid. The algorithm is run for Γ iterations until there are no more changes to the clusters. The clusters represented by centroids C^b and C^p are regions of potential tumor whereas C^w represents non-tumor. The blue and pink clusters are further investigated at a higher level of magnification for detailed classification. After K-means, the data is passed on to the next step, with the following values populated for each pixel P^i. Map M contains { $m_{p_1}, m_{p_2}...$} and each m_{p_i} *=(pixel-location, color value, label)*

2.3. *Otsu multi-level threshold segmentation*

A 2-level threshold segmentation is used in the next step. A window of 512 x 512 is considered and the color image is converted to 24 bit grayscale image, with more weight to blue channel. This is due to the fact that tumor regions have more blue channel values than non-tumor regions. The gray scale values for two level threshold are represented as [1,2,...t] and [t+1,....L] respectively and the weighted class variance is calculated as

$$\sigma_w^2(t) \ = \ q_1(t)\sigma_1^2(t) \ + \ q_2(t)\sigma_2^2(t) \tag{4}$$

where

$$q_1(t) \ = \sum_{i=1}^{t} P(i);\ q_2(t) \ = \sum_{i=t+1}^{L} P(i) \tag{5}$$

are the class probabilities and the intra-class variance is given by

$$\mu_1(t) = \sum_{i=1}^{t} \frac{i * P(i)}{q_1(t)} \tag{6}$$

Now the image contains two classes of pixels following a bi-modal histogram. We calculate the optimum threshold separating the two classes so that their combined spread (intra-class variance) is minimal. Otsu thresholding creates a red-blue color map where red signifies the non-viable tumor pixel and blue is the viable tumor or coagulative necrosis pixel. The data from this stage is exported by updating the values in map M as m_{p_i}=(pixel, red/blue value, original color value, label)

2.4. *Calculating clusters*

This stage calculates blue clusters and their properties. We run Flood-fill algorithm to identify boundaries and compute clusters. Viable tumor and coagulative necrosis always are in the blue region and contain cellular structures within them of non-uniform circularity. Along with each cluster, the size of the cluster in terms of area a, circularity c, and average color of cluster are calculated. Clusters that are less than minimum cluster size (50 pixels) and greater than maximum cluster size (300 pixels) are discarded as they represent false positives. The output of this stage is a map M' (Cluster label, Start Point, Centroid, Circularity, Area, List of Points, Color) and an image of clusters with red borders.

3. Results

The application was created using Openslide-Java for processing SVS images, ImageJ and Java Advanced Imaging for basic image processing tasks and C# .NET to perform Otsu segmentation and Flood-fill. The dataset included 120 images of 1160 x 640 resolution and all data samples were manually analyzed and classified by pathologists at UT Southwestern. The output of the application was compared with the classified images for verification and was validated by the team at UT Southwestern.

The choice of parameters such as window size in Otsu step (512*512), minimum and maximum cluster sizes (50 pixels and 300 pixels) etc. are selected based on empirical values for which the accuracy is maximized. Each figure (Figure 3 - Figure 6) consists of an original image I from the dataset, shown in (a), the output of Otsu method applied on I, (b) and clustering of cells by flood fill method,(c). Each cluster inside the red boundary in the images is defined in memory as a map data structure, Cluster(*Centroid, OriginalColor, Area, Perimeter, Circularity*).

3.1. *Viable Tumor*

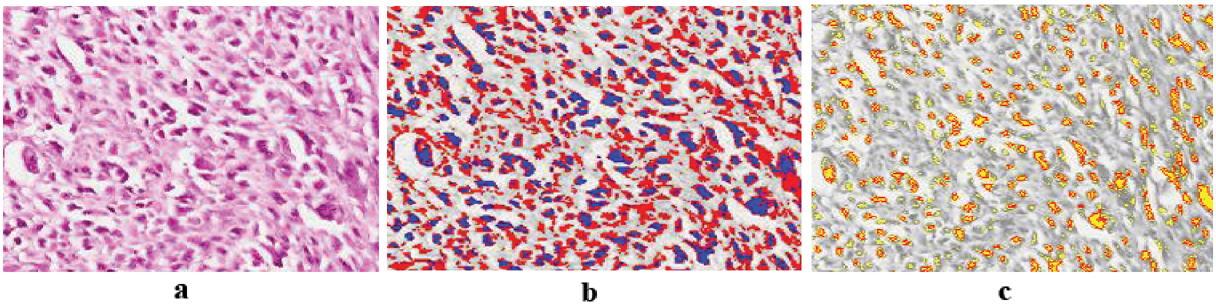


Fig. 3. Region: Viable tumor. **(a)** original image for viable tumor, **(b)** Otsu output showing more blue color, **(c)** Cell clustering using flood fill showing computed clusters.

Figure 3(a) shows that viable tumor has dense nuclei with more blue color. Otsu segmentation captures the nuclei with high accuracy represented by blue regions, as seen in 3(b). The percentage of blue pixels higher than the percentage of red pixels is a significant indicator for classifying this region into class Ψ_1. Clustered cellular information from Flood-fill 3(c) shows that there are more cells in the viable tumor region than the others. (see Figure 8(a)).

3.2. *Coagulative Necrosis*

Figure 4(a) shows coagulative necrosis containing cells with disintegrated nuclear matter, which makes the image appear brighter than viable tumor. Otsu segmentation step 4(b) yields higher percentage of blue pixels than red pixels, which is a key parameter in deciding regions belonging to class Ψ_1. Cluster properties in 4(c) show that the cell clusters are less dense and are distant from each other.

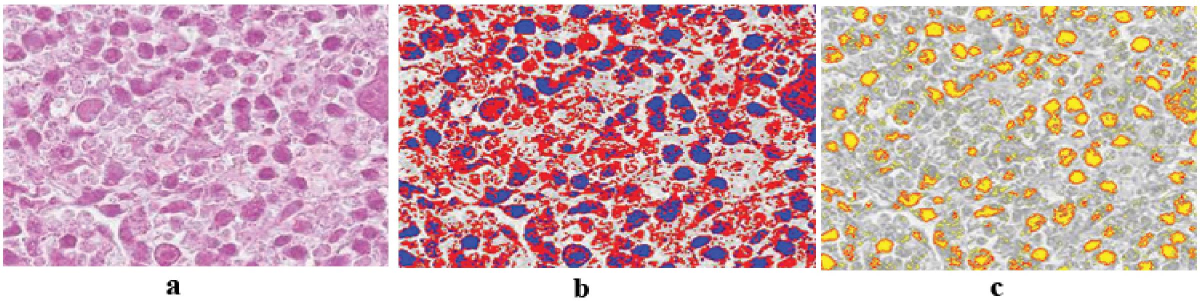


Fig. 4. Region: Coagulative necrosis. (**a**) original image for coagulative necrosis. (**b**) output of Otsu showing more blue than red, (**c**) Cell clustering using flood fill similar to viable tumor

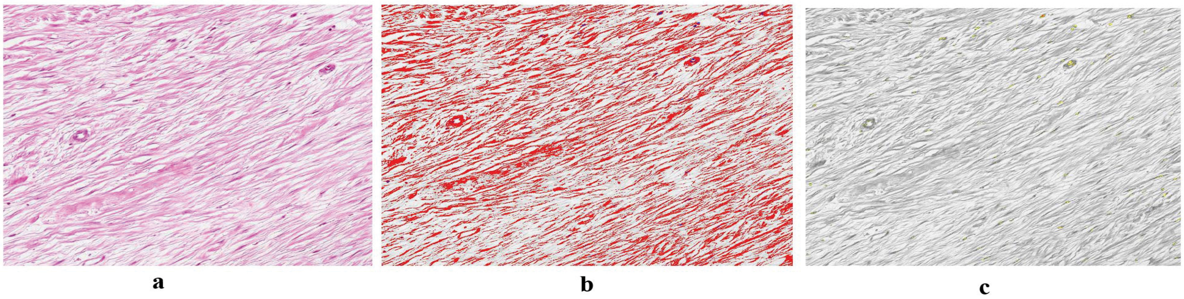


Fig. 5. Region: Fibrosis. (**a**) original image for fibrosis, (**b**) shows Otsu output for Fibrosis where there is a higher percentage of red pixels than blue, (**c**) cell clustering using flood fill, showing presence of fewer cells

3.3. *Fibrosis*

Figure 5(a) shows fibrosis region represented by strand like structures and absence of cells and nuclei. Due to this characterstic, Otsu in 5(b) produces higher percentage of red than blue pixels. This result distinguishes fibrosis from images in class Ψ_1. Flood-fill on this output, seen in 5(c), produces lesser number of clusters compared to viable tumor and coagulative necrosis.

3.4. *Osteoid*

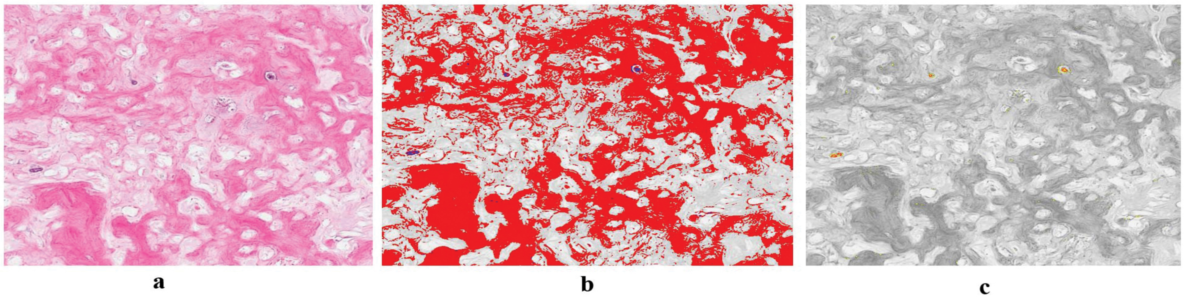


Fig. 6. Region: Osteoid. (**a**) original image for osteoid, (**b**) Otsu output for osteoid characterized by higher percentage of red than blue pixels. (**c**) Cell clustering flood fill output marked by absence of cells.

Figure 6(a) shows osteoid from the dataset, characterized by pink regions, background stroma and absence of cells and nuclei similar to fibrosis. Running Otsu on this image produces 6(b), with high percentage of red pixels due to absence of cells. This result makes osteoid to be grouped in Ψ_2 distinguishing them from images in class Ψ_1. Since osteoid is an extracellular protein matrix, it remains after tumor cells have undergone necrosis. However, there maybe interspersed cells found in the matrix. Flood-fill on this output, shown in 6(c), captures scattered cells that are less dense, unlike viable tumor and coagulative necrosis.

3.5. *Data Interpretation*

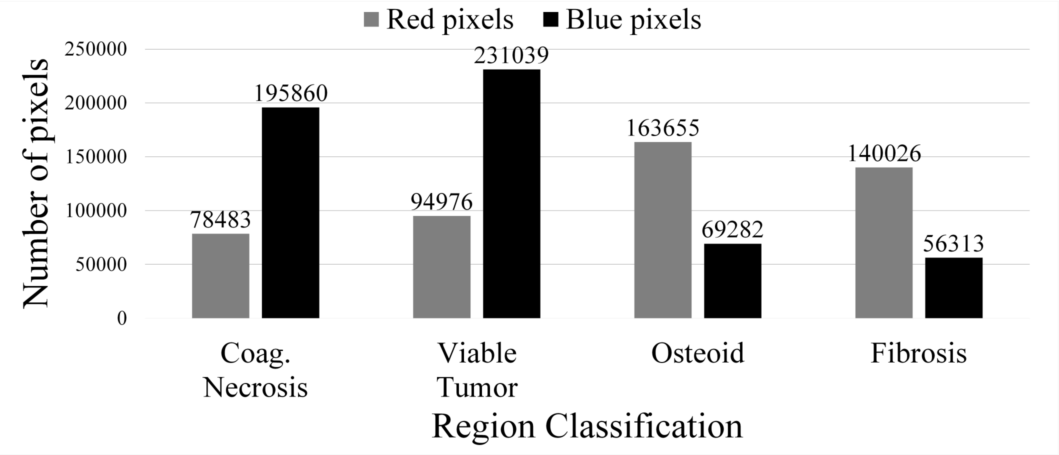


Fig. 7. Region wise average red and blue pixel count

A plot of average pixel count for classification regions shows that viable tumor and coagulative necrosis regions have more blue pixels, while fibrosis and osteoid regions have more red pixels. This result from Otsu step divides the image into prominent blue and red regions (see Figure 7). The regions that get classified under Ψ_1 have more cells than the regions under Ψ_2, and therefore have more blue pixels than red. Thus, images with viable tumor and coagulative necrosis regions can be classified into Ψ_1, while fibrosis and osteoid can be classified into Ψ_2.

A further analysis on average cell counts shows that viable tumor has 1702 cell clusters, while coagulative necrosis has 850 cells (see Figure 8(a)). This further distinguishes images in Ψ_1 into viable tumor and coagulative necrosis more accurately. We calculated the average density of cells in a 32x32 window as shown in Figure 8(b). It is observed that viable tumor has a cellular density of 2.4 while coagulative necrosis has 1.17. This important characteristic differentiates viable tumor from coagulative necrosis. The findings conclude that viable tumor is more dense and has closely aggregated cells than coagulative necrosis, the result of which has been used in classification.

It can be seen that fibrosis and osteoid regions have low cell clusters and high background stroma, hence concurring with the previous findings that these segmented images have less blue and more red pixels. Thus, the images in Ψ_2, that were identified as fibrosis and osteoid, have been categorized as non-viable tumor. Furthermore, in class Ψ_1, the cellular density distinguishes viable tumor from coagulative necrosis, which can be used to classify coagulative necrosis under non-viable tumor.

The accuracy of the method has been measured and is found in Table 1.

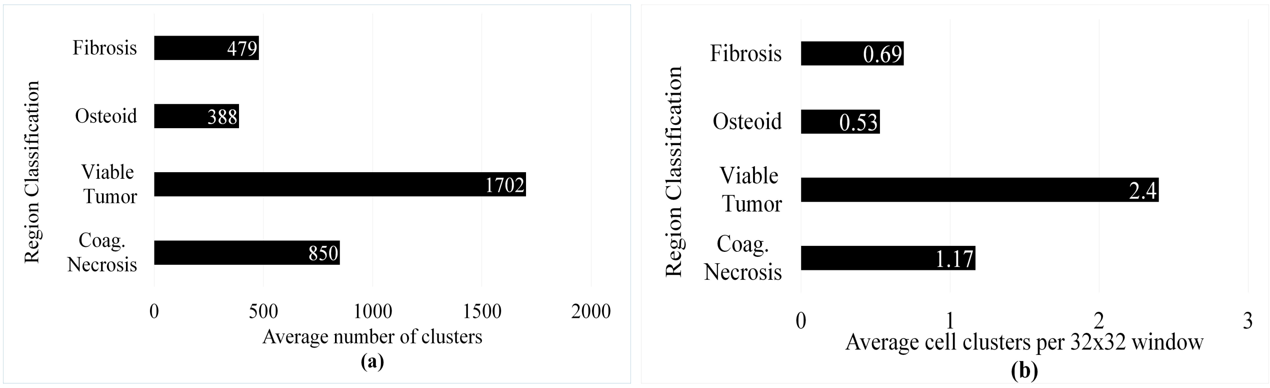


Fig. 8. (a) Region wise cellular count.(b) Region wise average cellular density count per 32x32 window.

Table 1. Quantitative metric comparison of classification regions

Region Type	Quantitative Metrics (Average)				
	red pixels count	blue pixels count	cell clusters	cell density (32x32 window)	Classification accuracy(%)
Viable Tumor	94976	231039	1702	2.4	100
Coagulative Necrosis	78483	195860	850	1.17	100
Osteoid	163655	69282	388	0.53	93
Fibrosis	140026	56313	479	0.69	89

4. Limitations and Future Improvements

The approach presented in this paper is limited to Osteosarcoma tumor identification. The current method works well in the given sampled datasets. We propose to extend it to all images in the dataset by incorporating contextual information that yield additional data. RGB color channels used in the experiments are affected by color variance in images and hence we would replace them with LAB colorspace. We plan to identify more relevant features from the images and subsequently use machine learning algorithms on them to improve classification accuracy.

Acknowledgments

This research was partially supported by NSF award IIP1439718 and CPRIT award RP150164. We would like to thank Dr.Lan Ma, University of Maryland and Dr.Riccardo Ziraldo, University of Texas at Dallas for their helpful discussions. We also would like to thank John-Paul Bach and Sammy Glick from UT Southwestern Medical Center for their help with datasets.

References

1. J. D. Webster, A. M. Michalowski, J. E. Dwyer, K. N. Corps, B.-R. Wei, T. Juopperi, S. B. Hoover, R. M. Simpson *et al.*, *Journal of pathology informatics* **3**, p. 18 (2012).
2. J. I. Zhongwu Wang, John R. Jensen, *Environmental Modelling and Software* **25**, 1149 (October 2010).
3. R. Harrabi and E. B. Braiek, *EURASIP Journal on Image and Video Processing* **2012**, 1 (2012).
4. G. Ottaviani and N. Jaffe, *The Epidemiology of Osteosarcoma*, in *Pediatric and Adolescent Osteosarcoma*, eds. N. Jaffe, S. O. Bruland and S. Bielack (Springer US, Boston, MA, 2010), Boston, MA, pp. 3–13.
5. A. H. Fischer, K. A. Jacobson, J. Rose and R. Zeller, *Cold Spring Harbor Protocols* **2008**, pdb (2008).
6. Aperio svs tiff format `http://openslide.org/formats/aperio/`.
7. M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot and B. Yener, *IEEE Reviews in Biomedical Engineering* **2**, 147 (2009).
8. T. H. S. Sonal Kothari, John H Phan and M. D. Wang, *J Am Med Inform Assoc.* **20**, 1099 (November 2013).
9. A. N. Basavanhally, S. Ganesan, S. Agner, J. P. Monaco, M. D. Feldman, J. E. Tomaszewski, G. Bhanot and A. Madabhushi, *IEEE Transactions on Biomedical Engineering* **57**, 642 (March 2010).
10. T. J. Fuchs and J. M. Buhmann, *Computerized Medical Imaging and Graphics* **35**, 515 (2011).
11. K.-H. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. Ré, D. L. Rubin and M. Snyder, *Nature Communications* **7** (2016).
12. K. L. Weind, C. F. Maier, B. K. Rutt and M. Moussa, Invasive carcinomas and fibroadenomas of the breast: comparison of microvessel distributions–implications for imaging modalities. *Radiology* **208**1998. PMID: 9680579.
13. T. Blaschke, G. J. Hay, M. Kelly, S. Lang, P. Hofmann, E. Addink, R. Q. Feitosa, F. van der Meer, H. van der Werff, F. van Coillie *et al.*, *ISPRS Journal of Photogrammetry and Remote Sensing* **87**, 180 (2014).
14. S. P. Brumby, N. R. Harvey, S. J. Perkins, R. B. Porter, J. J. Szymanski, J. P. Theiler and J. J. Bloch, Genetic algorithm for combining new and existing image processing tools for multispectral imagery, in *AeroSense 2000*,
15. A. Goode, B. Gilbert, J. Harkes, D. Jukic, M. Satyanarayanan *et al.*, Openslide: A vendor-neutral software foundation for digital pathology *Journal of pathology informatics* **4** (Medknow Publications, 2013).
16. K. Martinez and J. Cupitt, Vips-a highly tuned image processing software architecture, in *IEEE International Conference on Image Processing 2005*, 2005.
17. F. Wang, T. W. Oh, C. Vergara-Niedermayr, T. Kurc and J. Saltz, Managing and querying whole slide images, in *SPIE Medical Imaging*, 2012.
18. M. D. Abramoff, P. J. Magalhaes and S. J. Ram, *Biophotonics international* **11**, 36 (2004).
19. T. R. Jones, I. H. Kang, D. B. Wheeler, R. A. Lindquist, A. Papallo, D. M. Sabatini, P. Golland and A. E. Carpenter, *BMC bioinformatics* **9**, p. 1 (2008).

MISSING DATA IMPUTATION IN THE ELECTRONIC HEALTH RECORD USING DEEPLY LEARNED AUTOENCODERS[*]

BRETT K. BEAULIEU-JONES

Genomics and Computational Biology Graduate Group, Computational Genetics Lab, Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia PA, 19104
Email: brettbe@med.upenn.edu

JASON H. MOORE

Computational Genetics Lab, Institute for Biomedical Informatics, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia PA, 19104
Email: jhmoore@exchange.upenn.edu

THE POOLED RESOURCE OPEN-ACCESS ALS CLINICAL TRIALS CONSORTIUM[†]

Electronic health records (EHRs) have become a vital source of patient outcome data but the widespread prevalence of missing data presents a major challenge. Different causes of missing data in the EHR data may introduce unintentional bias. Here, we compare the effectiveness of popular multiple imputation strategies with a deeply learned autoencoder using the Pooled Resource Open-Access ALS Clinical Trials Database (PRO-ACT). To evaluate performance, we examined imputation accuracy for known values simulated to be either missing completely at random or missing not at random. We also compared ALS disease progression prediction across different imputation models. Autoencoders showed strong performance for imputation accuracy and contributed to the strongest disease progression predictor. Finally, we show that despite clinical heterogeneity, ALS disease progression appears homogenous with time from onset being the most important predictor.

[*] This work is supported by a Commonwealth Universal Research Enhancement (CURE) Program grant from the Pennsylvania Department of Health.

[†] Data used in the preparation of this article were obtained from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Database. As such, the following organizations and individuals within the PRO-ACT Consortium contributed to the design and implementation of the PRO-ACT Database and/or provided data, but did not participate in the analysis of the data or the writing of this report: Neurological Clinical Research Institute, MGH, Northeast ALS Consoritium, Novartis, Prize4Life, Regeneron Pharmaceuticals, Sanofi, Teva Pharmaceutical Industries, Ltd.

1. Introduction

1.1. *Background*

Electronic health records (EHRs) are a core resource in genetic, epidemiological and clinical research providing phenotypic, patient progression and outcome data to researchers. Missing data presents major challenges to research by reducing viable sample size and introducing potential biases through patient selection or imputation[1,2].

Missing data is widely prevalent in the EHR for several reasons. EHRs are designed and optimized for clinical and billing purposes meaning data useful to research may not be recorded[2]. Outside of the design of EHRs, the reality of the clinic results in missing data. For example, clinicians must consider financial burden in ordering lab tests for patients and issue the minimum amount of testing and diagnostics to effectively treat their patients[3].

The various reasons data may be missing create different types of missing data: missing completely at random, missing at random and missing not at random[1,4,5]. This work focuses on data missing completely at random and data missing not at random. Data missing completely at random indicates that there is no systematic determination of whether a value is missing or present. The likelihood of response is independent of the data and any latent factors. Data missing not at random occurs when data is missing due to either observed values in the data or unobserved latent values. An example within the EHR would occur if a lab test is only issued based on a clinicians observation of the patient. Whether or not the test is issued provides insight into the patient's status.

Non-imputation approaches often exclude data from the analysis to allow for downstream analysis. One approach, complete-case analysis, throws out records with missing data. Within the EHR this would severely limit sample size, in addition if incomplete records have a systematic difference from complete records unintentional bias can be introduced[6]. As computational resources have increased, computationally complex imputation techniques have become feasible and are growing in popularity[1]. Computationally intensive techniques such as Singular Value Decomposition (SVD) based methods and weighted K-nearest neighbors (KNN) methods have joined less complex methods like mean and median imputation. Both SVD and KNN-based methods have been shown to perform effectively in microarray imputation[7]. Popular multiple imputation methods show particular challenges with data that are not missing at random[1].

Autoencoders are a variation of artificial neural networks that learn a distributed representation of their input[8]. They learn parameters to transform the data to a hidden layer and then reconstruct the original input. By using a hidden layer smaller than the number of input features, or "bottle-neck" layer the autoencoder is forced to learn the most important patterns in the data[9]. To prevent overreliance on specific features two techniques are commonly used. In a denoising autoencoder, noise is added to corrupt a portion of the inputs[9,10]. Alternatively, a technique called dropout in which random units and connections are removed from the network forcing it to learn generalizations[11]. Autoencoders were shown to generate useful higher representations in both simulated and real EHR data. Because autoencoders learn by reconstructing the original input from a corrupted version, imputation is a natural extension[12,13].

1.2. *ALS and the Pooled Resource Open-access Clinical Trials*

We evaluate each of the imputation methods on the ALS Pooled Resource Open-access Clinical Trials (PRO-ACT). Pooled clinical trial datasets present an ideal option for evaluating EHR imputation strategies because they include patients from differing environments with potential systematic biases. In addition, clinical trials represent the gold standard for data collection making it possible to spike-in missing data while maintaining enough signal to evaluate imputation techniques.

Prize4Life and the Neurological Clinical Research Institute (NCRI) at Massachusetts General Hospital created the Pooled Resource Open-Access ALS Clincial Trials (PRO-ACT) platform with funding from the ALS Therapy Alliance in and in partnership with the Northeast ALS Consortium. The PRO-ACT project was designed to empower translational ALS research and includes data from 23 clinical trials and 10,723 patients. In this work, we use the subset of 1,824 patients included in the Prize4Life challenges[14].

ALS is a progressive neurodegenerative disorder affecting both the upper and lower motor neurons causing muscle weakness, paralysis and leading to death[14]. ALS patients typically survive only 3 to 5 years from disease onset and show large degrees of clinical heterogeneity[15–18].

A common measure used to monitor an ALS patient's condition is the ALS functional rating scale (ALSFRS)[19,20]. The ALSFRS consists of 10 tests scored from 0-4 assessing patients' self-sufficiency in categories including: feeding, grooming, ambulation and communication. The change over time, or slope, is commonly used as a statistic to represent ALS progression.

2. Methods

We compare and evaluate a variety of methods to impute missing data in the EHR. We spiked-in missing data to the PRO-ACT dataset, and evaluated each approach's performance imputing known data. We also evaluated prediction accuracy using each of the imputation methods on the ALSFRS. Each of these is described in detail below and all analysis was run using freely available open source library packages, DAPS[12], FancyImpute[21], Keras[22] and Scikit-learn[23].

2.1. *Data preparation and standardization*

The PRO-ACT dataset includes patient demographic data, family history, concomitant medications, vital sign measurements, laboratory results, and patient history (disease onset etc.). PRO-ACT performed an initial data cleaning and quality assurance process. This process included extracting quantitative variables, merging laboratory tests with different names across trials, removable of indecipherable records and converting units. After processing the PRO-ACT dataset includes only quantitative values (continuous, binary, ordinal and categorical).

Our analysis encoded categorical variables using Sci-kit learn's OneHotEncoder[23]. Temporal or repeated measurements were encoded as the mean, minimum, maximum, count, standard deviation and slope across all measurements, creating 572 features for each samples. Additional measurements were standardized across scales (i.e. inches to cm). Non-numeric values in numeric measurements were coerced to numeric values. Where coercion failed they were replaced by NaN.

Input features were normalized and scaled to be between 0 and 1, with missing features remaining as NaN.

2.2. *Imputation Strategies*

2.2.1. *Imputing missing data with Autoencoders*

We constructed an autoencoder with a modified binary cross entropy cost between the reconstructed layer z and the input data x to better handle missing data as in Beaulieu-Jones and Greene (2016) (Formula 1)[12]. The modified function takes into account missing data, with m representing a "missingness" vector; m has a value of 1 where the data is present and 0 when the data is missing. By multiplying by m and dividing by the count of present features (sum of m) the result represents the average cost per present feature. The weights and biases of the autoencoder are trained only on present features and imputation does not need to be performed prior to training the autoencoder.

$$cost = -\sum_{k=1}^{d}[x_k \log(z_k)\ m_k + (1-x_k)\log(1-z_k)\, m_k] \,/\, count(m) \qquad \text{(Formula 1)}$$

With the exception of the modified cost function autoencoders were trained as described by Vincent et al. with a 100 training epoch patience[10]. If a new minimum cost was not reached in 100 epochs, training was stopped. The autoencoder with dropout was implemented using the FancyImpute[21] and Keras[22] libraries with a Theano[24,25] backend.

We performed a parameter sweep to determine the hyperparameters for the autoencoder. In the sweep we included autoencoders of one to three hidden layers and each combination of 2, 4, 10, 100, 200, 500 and 1000 (over-complete representation) hidden nodes per layer. Autoencoders with two hidden layers made up of 500 nodes each are shown for all comparisons (Figure 1). Dropout levels of 5, 10, 20, 30, 40 and 50% were evaluated with 20% being shown for all comparisons.

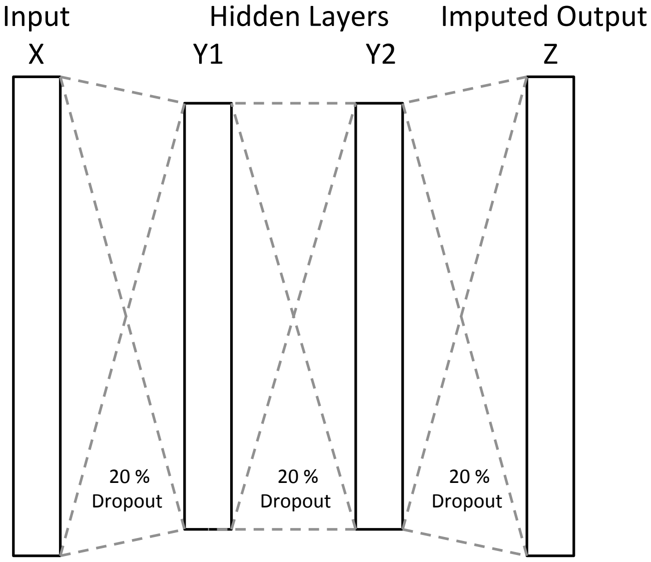


Figure 1. Schematic structure of the autoencoder used for evaluations, with two hidden layers and 20% dropout between each layer.

Binary cross entropy was used for training because it tends to be a better evaluator of quality when training neural networks[9,10,26,27]. We use a root mean squared error for comparison to other methods to prevent a bias in favor of autoencoders, as most other methods are not trained with cross entropy.

2.2.2. *Comparative imputation strategies*

We used the FancyImpute[21] libraries implementations for each of the other imputation strategies: 1) IterativeSVD, matrix completion by low rank singular value decomposition based on SVDImpute[7], 2) K-nearest neighbors imputation (KNNimpute), matrix completion by choosing the mean values of the K closest samples for features where both samples are present 3) SoftImpute[28], matrix completion by iterative replacement of missing values with values from a soft-thresholded singular value decomposition, 4) column mean filling and 5) column median filling. The standard implementations of the remaining algorithms in the FancyImpute library, MICE, Matrix Factorization and Nuclear Norm Minimization are known to be slow on large matrices and were impractically slow on this dataset[29–33].

We performed a parameter sweep for SVDimpute analyzing ranks of 5, 10, 20, 40 and 80. Ranks of 40 showed the strongest performance with this dataset and are shown for all comparisons. The parameter sweep for KNNimpute included 1, 3, 5, 7, 15 and 30 neighbors, k of 7 showed the strongest performance of the parameter sweep and is used for all comparisons.

2.3. *Missing Completely at Random Imputation Evaluation*

To evaluate imputation accuracy in a missing completely at random environment we performed trials replacing 10, 20, 30, 40 and 50% of known features at random with NaN. We performed each imputation strategy on the data with spiked-in missingness and evaluated the root mean squared error between the imputed estimates and the original data. We performed five trials for each amount of spiked-in data (Figure 2A). Performance was evaluated using the root mean squared error between the known value before spiking in missingness and the imputed value.

2.4. *Missing Not at Random Imputation Evaluation*

To perform a basic imputation simulation where data was missing not at random, varying percentages (10, 20, 30, 40, and 50%) of features were chosen at random. Half of the highest or lowest (randomly selected) quartile of values was replaced by NaN at random. Each imputation strategy was evaluated on five independent spike-in trails. Performance was evaluated using the root mean squared error between the imputed values and original values. This type of imputation could occur when the highest or lowest values represent the normal range and the clinician is able to determine a patient is normal through other factors. Alternatively the extreme values could represent a clear result where an additional is not needed to determine the result. Performance was evaluated using the root mean squared error between the known value before spiking in missingness and the predicted value.

2.5. *Progression Prediction Evaluation*

To predict disease progression as represented by the ALSFRS score slope, we first imputed the missing data using column mean averaging, column median averaging, SVDImpute, SoftImpute, KNNimpute, and an autoencoder with dropout. For prediction purposes we excluded all ALSFRS score and Forced Vital Capacity-related features.

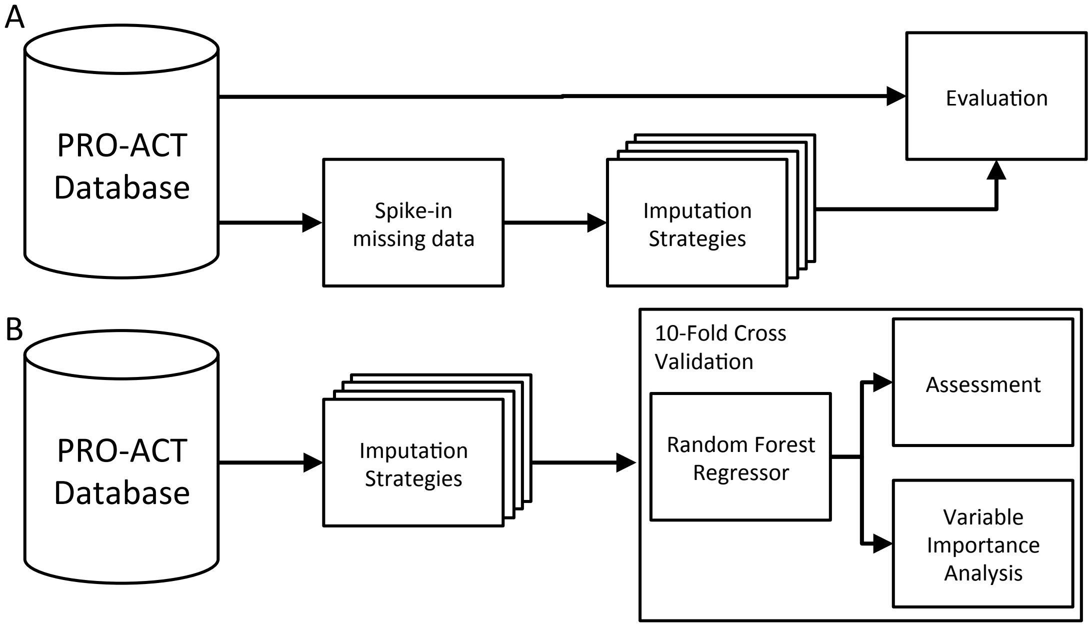


Figure 2. Evaluation outline **(a)** Imputation Evaluation. PRO-ACT patient data of 10,723 subjects has known data masked with spiked in missing data. Imputation strategies are performed in parallel and the RMSE is calculated between the masked input data and each strategies imputations. **(b)** Progression Prediction. PRO-ACT patients are imputed using each strategy. Ten-fold cross validation of a random forest regressor is performed on imputed patients.

We then used the scikit-learn implementation of a random forest regressor[23] to predict the ALSFRS score slope. The random forest regressor was chosen because four of the top six teams in the DREAM-Phil Bowen ALS Prediction Prize4Life challenge used variants of random forest regressors[14]. We also compare a random forest regressor modified to predict progression from the raw data without imputation[34]. Ten-fold cross validation was performed and the root mean squared error between the predicted slope and actual slope was calculated. We then extracted the top 10 most important features used in the trained model for analysis (Figure 2B).

3. Results

Most patients were missing approximately half of the features we extracted from the EHR (3A). The pooled aspect of the PRO-ACT data is particularly evident in the distribution of missing features as different clinical trials collected different amounts of data. Features tended to be observed in either less than 25% or in greater than 75% of patients (Figure 3B). Lab tests in particular demonstrated high variability of missingness among patients, with many present in small numbers of patients. It is impossible to determine the level of each type of missing data that exists, but it is clear that at least some of data missing is due to clinical factors (trial group etc.). The most complete features are demographics and family history information, information likely collected before entry into any of the clinical trials.

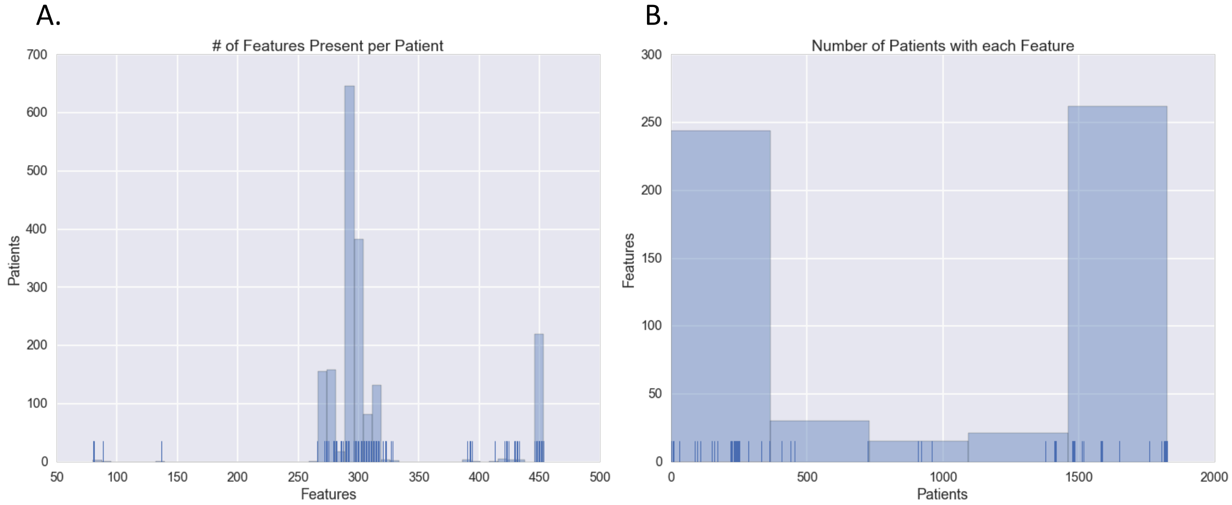


Figure 3. Histogram distribution and rug plot showing the number of patients each feature is present in. **(a)** The number of features each patient has. Ticks at the bottom indicate one patient with the count of features, bins indicate the number of patients in a range. **(b)** The number of patients having a recorded value for each feature. Ticks at the bottom indicate the number of patients a feature is present in, bins indicate the number of features in a range.

3.1. *Missing completely at random spike-in results*

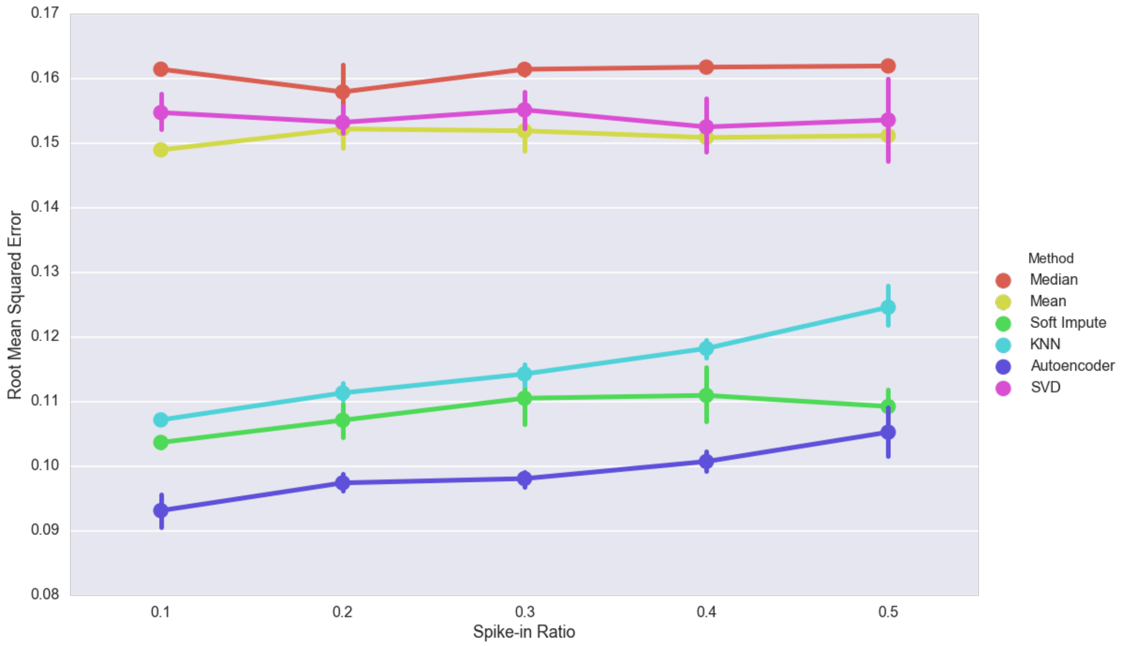


Figure 4. Effect of the amount of spiked-in missing data on imputation. Error bars indicate 5-fold cross validation score ranges.

Mean, Median and Singular Value Decomposition imputation perform poorly when data is missing completely at random. However, they do not appear to degrade as the spike-in ratio increase (Figure 4). This is not surprising for mean and median imputation because missing data is chosen completely at random and is unlikely to have a large effect on statistical averages. The autoencoder had the highest imputation performance despite increasing as the spike-in ratio increased.

3.2. *Not missing at random spike-in results*

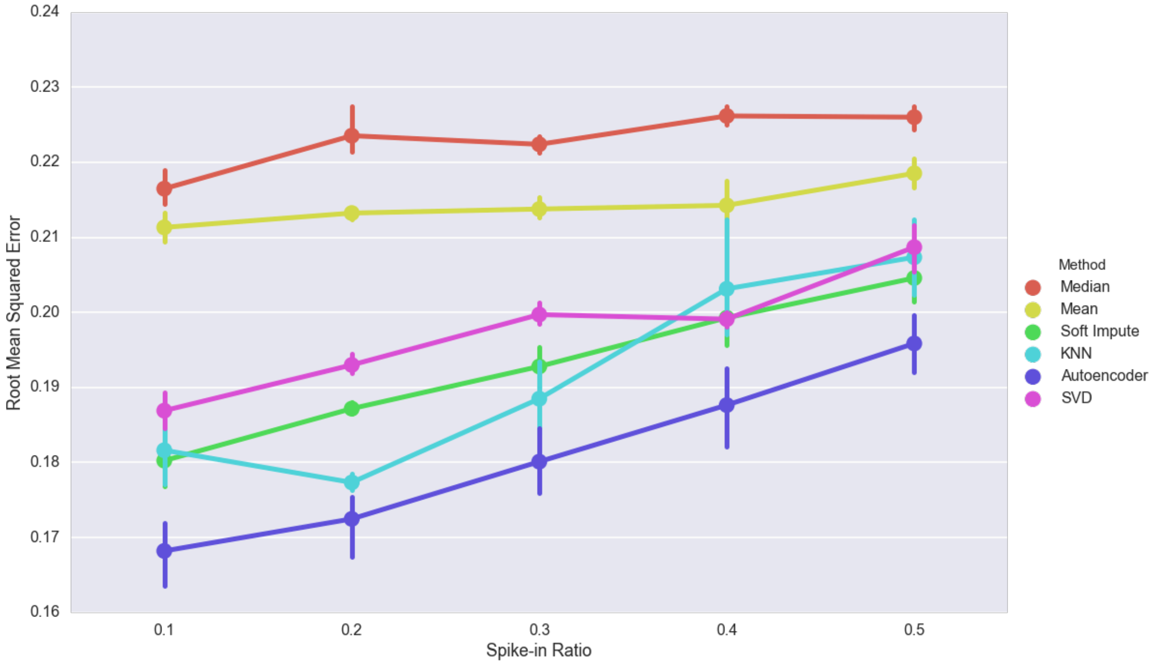


Figure 5. Effect of non-random spiked-in missing data on imputation (measured in root mean squared error). Autoencoder w/Dropout (2 layer 500 nodes each), SVD – SVDImpute with rank of 40, KNN - KNNimpute with 7 neighbors, Mean – Column Mean Averaging, Median – column median averaging, SI – SoftImpute.

The trends seen in the missing completely at random experiment largely repeat when the data is missing not at random. The autoencoder approach shows strong performance but is closely followed by the KNN, Softimpute and SVD approaches (Figure 5). KNN works by finding the k-nearest neighbors for shared values and taking the mean for the missing feature. Autoencoders work by learning the optimal network for reconstruction. Similar input values will have similar hidden node values. This similarity could explain the relatively even performance between the two methods. In addition to recognizing similar samples, autoencoders have been shown to perform well when there is dependency or correlation between variables[35]; this is the scenario when data is missing not at random. When spike-in ratios increase to high levels the methods begin to converge to the performance of mean and median imputation. This is likely because too much of the signal is lost as missing data to learn the correlation structure.

3.3. *ALS disease progression*

Imputation strategy has a modest but statistically significant impact on the root mean squared error of ALS disease progression prediction, but the autoencoder approach is the strongest performing

(Figure 6). Despite showing poor performance in the imputation accuracy exercises Singular Value Decomposition does approximately as well as k-nearest neighbors and SoftImpute in this experiment. A random forest regressor applied to the raw data is the worst performing, but is not significantly worse than any of the methods other than the Autoencoder. In terms of ALS disease progression, imputation does not appear to have a large effect on prediction, but can be vital to allow the use of other algorithms (prediction, clustering etc.) without modification.

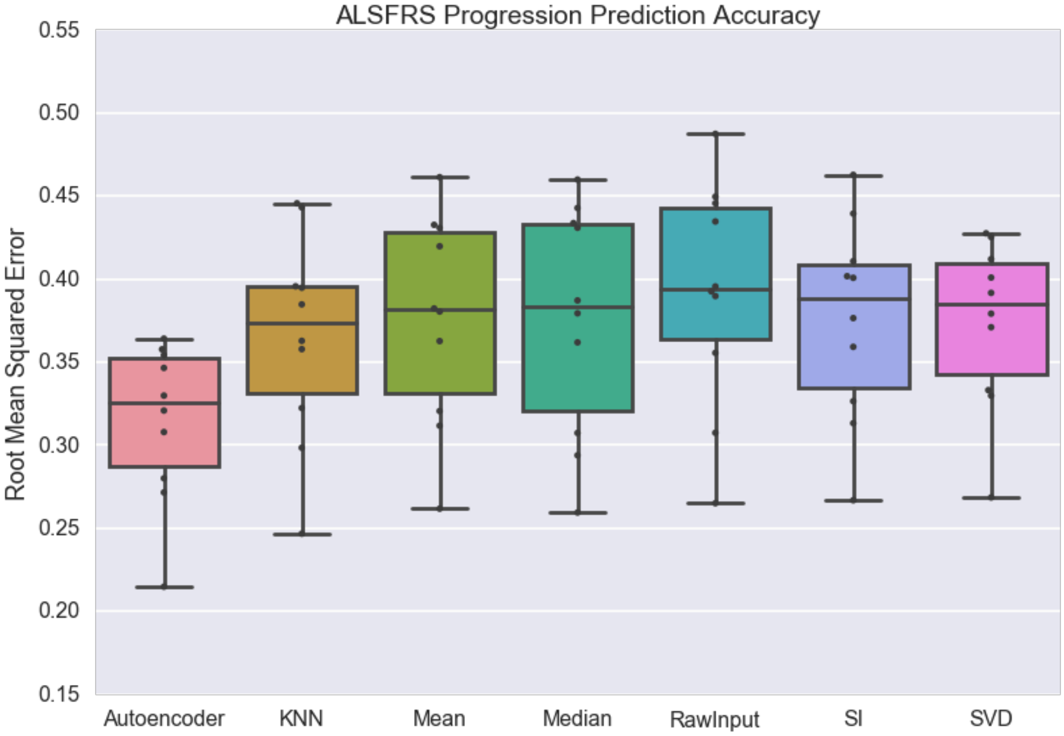


Figure 6. ALS Functional Rating Scale prediction accuracy shown for an autoencoder, k-nearest neighbors, mean averaging, median averaging, the raw input including missing values, soft impute and singular value decomposition. The box indicates inner quartiles with the line representing the median; the whiskers indicate outer quartiles excluding outliers.

3.4. *ALS progression predictive indicat*

Nine out of the top ten most important features in the autoencoder-imputed random forest regressor were among the top fifteen identified in the DREAM Prediction challenge (Figure 7A). The amount of time using Riluzole was not among the top fifteen previously identified. Riluzole is the only FDA approved medication for ALS treatment but it is believed to have a limited effect on survival[36–38]. The finding that Riluzole is protective of ALS slope indicates some level of efficacy.

Of the top ten most important features, five are missing in more than 50% of patients in the data set. This is a possible explanation for the improvement shown by Autoencoders, SVDimpute and KNNimpute over mean imputation.

By far the most important feature for prediction is the time from onset and several of the most important features are highly correlated with time from onset. ALSFRS slopes resemble a normal distribution (Figure 7B). When including the entire PRO-ACT dataset, the Kolmogorov-Smirnov test score is 0.05 for patients with negative slopes. This indicates the progression of the disease is similar to a truncated normal distribution. We exclude positive slopes because ALS patients do not typically get better, and signs of doing so are likely the result of measurement error. Despite presenting in clinically heterogeneous manners, ALS progression as defined by the ALSFRS appears to be largely homogenous. Patients fall within a relatively normal distribution and have increasingly negative slopes the longer they ALS.

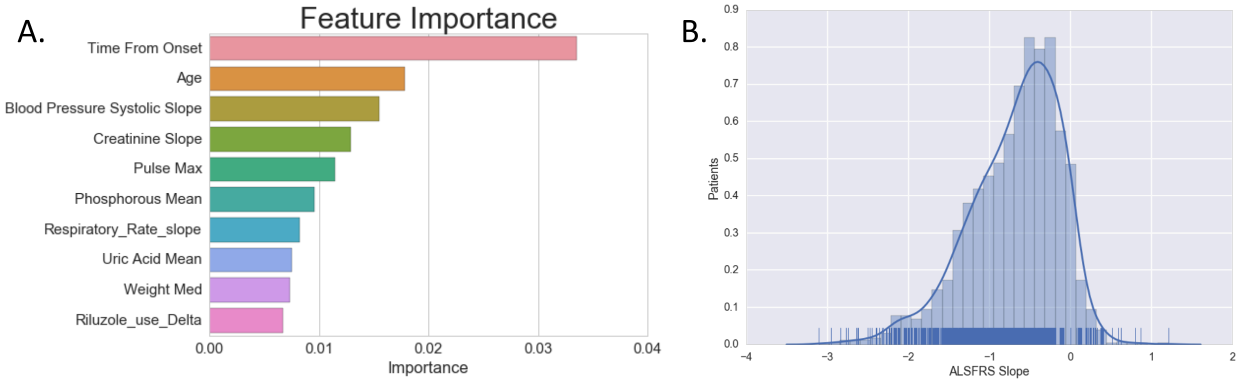


Figure 7. Prediction feature importance. **(a)** Importance levels of the top 10 features to the random forest regressor with autoencoder imputed data. **(b)** Histogram distribution of patient ALSFRS slope levels.

4. Discussion and Conclusions

In this study, we compared the performance of an autoencoder approach with popular imputation techniques in ALS EHR data. A multi-layer autoencoder with dropout showed robust imputation performance across a variety of spiked-in missing data experiments designed to be both completely at random and not at random. Furthermore, we found that imputation accuracy may not strictly correlate with predictive performance but the most accurate imputer provided the most accurate predictor. The importance of imputation is demonstrated by five of the top ten most important features for prediction being missing in more than 50% of patients.

Increased deterioration of imputation performance for KNNimpute and SVDimpute with increased missing data is at odds with previous research of imputation in microarrays[7]. Possible explanations include either reaching a threshold of missing data where the burden is too high for these methods to accurately impute or that a confounding systematic bias is introduced from the different clinical trials.

This work is a promising first step in utilizing deep learning techniques for missing data imputation in the EHR but challenges remain. Autoencoders are computationally intensive, but less so than imputation techniques like MICE, Matrix Factorization and Nuclear Norm Minimization. With GPU resources, autoencoders train in similar amounts of time to both KNN and SVD methods for these clinical trials. As data increases, autoencoder training time increases linearly in line with the number of samples. Methods like KNN require computing a distance matrix, which increases in exponential time. In addition, further examination is necessary to determine whether the strong performance shown by autoencoders is a result of the structure of this pooled clinical trial dataset. The subset of 1,800 patients is relatively small and methods may differ in performance increases with more patients.

This work offers promising results but has several limitations especially because it specifically analyzes pre-processed pooled clinical trial data. Clinical trials have more complete and cleaner data than raw EHR. Follow up work should be performed with other diseases and in the general patient population. These methods have also only been evaluated for quantitative values; in raw EHR data there will be an additional extraction step for raw text and qualitative observations that was not necessary due to PRO-ACT's preprocessing.

Additional future work will be concentrated on developing tools to better understand and interpret the structure of the trained autoencoder networks. We anticipate being able to recognize patterns in the trained weights to see correlation between input features. Understanding correlation will empower new clustering and visualization opportunities. Spike-in evaluations can provide a supervised context to otherwise unsupervised learning problems; further analysis should be performed on the higher-level learned features in the hidden layers of the autoencoders. We suspect these features may be useful in patient outcome classification and regression problems.

5. Acknowledgments

This work is supported by the Commonwealth Universal Research Enhancement (CURE) Program grant from the Pennsylvania Department of Health to JHM. The authors would like to thank Dr. Casey S Greene for helpful discussions. The authors acknowledge the support of the NVIDIA Corporation for the donation of a TitanX GPU used for this research.

References

1. Sterne JJ a C, White IRI, Carlin JJB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338(July):b2393. doi:10.1136/bmj.b2393.
2. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. *EGEMS (Washington, DC)*. 2013;1(3):1035. doi:10.13063/2327-9214.1035.
3. McClatchey KD. *Clinical Laboratory Medicine*. Lippincott Williams & Wilkins; 2002.
4. Little R, Rubin D. *Statistical Analysis with Missing Data*. John Wiley & Sons; 2014.
5. Marlin B. Missing data problems in machine learning. 2008. http://www-devel.cs.ubc.ca/~bmarlin/research/phd_thesis/marlin-phd-thesis.pdf. Accessed August 7, 2016.
6. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/hierarchical Models*.; 2006. https://books.google.com/books?hl=en&lr=&id=c9xLKzZWoZ4C&oi=fnd&pg=PR17&dq=data+analysis+using+regression+and+multilevel/hierarchical+models&ots=baT3R3Mnng&sig=KpLzVOFtUseaK8_IhUfPLM2Y7fU. Accessed August 10, 2016.
7. Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics*. 2001;17(6):520-525. doi:10.1093/bioinformatics/17.6.520.
8. Bengio Y. Learning Deep Architectures for AI. *Found Trends® Mach Learn*. 2009;2(1):1-127. doi:10.1561/2200000006.
9. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J Mach Learn Res*. 2010;11(3):3371-3408. doi:10.1111/1467-8535.00290.
10. Vincent P, Larochelle H, Bengio Y, Manzagol P-A. Extracting and composing robust features with denoising autoencoders. *Proc 25th Int Conf Mach Learn - ICML '08*. 2008:1096-1103. doi:10.1145/1390156.1390294.
11. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res*. 2014;15:1929-1958. doi:10.1214/12-AOS1000.
12. Beaulieu-Jones BK, Greene CS. Semi-Supervised Learning of the Electronic Health Record with Denoising

Autoencoders for Phenotype Stratification. *bioRxiv*. February 2016:39800. doi:10.1101/039800.

13. Miotto R, Li L, Kidd BA, et al. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep*. 2016;6:26094. doi:10.1038/srep26094.
14. Küffner R, Zach N, Norel R, et al. Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nat Biotechnol*. 2015;33(1):51-57. doi:10.1038/nbt.3051.
15. Kollewe K, Mauss U, Krampfl K, Petri S, Dengler R, Mohammadi B. ALSFRS-R score and its ratio: A useful predictor for ALS-progression. *J Neurol Sci*. 2008;275(1-2):69-73. doi:10.1016/j.jns.2008.07.016.
16. Beghi E, Mennini T, Bendotti C, et al. The heterogeneity of amyotrophic lateral sclerosis: a possible explanation of treatment failure. *Curr Med Chem*. 2007;14(30):3185-3200. http://www.ncbi.nlm.nih.gov/pubmed/18220753. Accessed August 7, 2016.
17. Sabatelli M, Conte A, Zollino M. Clinical and genetic heterogeneity of amyotrophic lateral sclerosis. *Clin Genet*. 2013;83(5):408-416. doi:10.1111/cge.12117.
18. Ravits JM, La Spada AR. ALS motor phenotype heterogeneity, focality, and spread: deconstructing motor neuron degeneration. *Neurology*. 2009;73(10):805-811. doi:10.1212/WNL.0b013e3181b6bbbd.
19. Cedarbaum JM, Stambler N. Performance of the amyotrophic lateral sclerosis functional rating scale (ALSFRS) in multicenter clinical trials. In: *Journal of the Neurological Sciences*. Vol 152. ; 1997. doi:10.1016/S0022-510X(97)00237-2.
20. Cedarbaum JM, Stambler N, Malta E, et al. The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function. *J Neurol Sci*. 1999;169(1-2):13-21. doi:10.1016/S0022-510X(99)00210-5.
21. Rubinsteyn A, Feldman S. fancyimpute: Version 0.0.9. March 2016. doi:10.5281/zenodo.47151.
22. Chollet F. Keras. *GitHub Repos*. 2015.
23. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. ... *Mach Learn* …. 2012;12:2825-2830. http://dl.acm.org/citation.cfm?id=2078195\nhttp://arxiv.org/abs/1201.0490.
24. Bastien F, Lamblin P, Pascanu R, et al. Theano: new features and speed improvements. *arXiv Prepr arXiv* …. 2012:1-10. http://arxiv.org/abs/1211.5590.
25. Bergstra J, Breuleux O, Bastien F, et al. Theano: a CPU and GPU math compiler in Python. In: *9th Python in Science Conference*. ; 2010:1-7. http://www-etud.iro.umontreal.ca/~wardefar/publications/theano_scipy2010.pdf.
26. Socher R, Pennington J, Huang E, Ng A. Semi-supervised recursive autoencoders for predicting sentiment distributions. *Proc*. 2011. http://dl.acm.org/citation.cfm?id=2145450. Accessed August 8, 2016.
27. Hinton G, Salakhutdinov R. Reducing the dimensionality of data with neural networks. *Science (80-)*. 2006. http://science.sciencemag.org/content/313/5786/504.short. Accessed August 8, 2016.
28. Mazumder R, Hastie T, Edu H, Tibshirani R, Edu T. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *J Mach Learn Res*. 2010;11:2287-2322.
29. Buuren S van. *Flexible Imputation of Missing Data*.; 2012. doi:10.1201/b11826.
30. Royston P. Multiple imputation of missing values: update of ice. *Stata J*. 2005. https://www.researchgate.net/profile/James_Cui2/publication/23780230_Buckley-James_method_for_analyzing_censored_data_with_an_application_to_a_cardiovascular_disease_and_an_HIVAIDS_study/links/53d5866d0cf228d363ea0b7a.pdf#page=59. Accessed August 10, 2016.
31. Kim J, Park H. Toward Faster Nonnegative Matrix Factorization: A New Algorithm and Comparisons.
32. Lin C-J. Projected Gradient Methods for Non-negative Matrix Factorization.
33. Hsieh C-J, Olsen PA. Nuclear Norm Minimization via Active Subspace Selection.
34. Breiman L, Cutler A. Random Forests. 2004. *URL http//stat-www berkeley edu/users/breiman/RandomForests/cc home htm*. 2014.
35. Nelwamondo F V., Mohamed S, Marwala T. Missing Data: A Comparison of Neural Network and Expectation Maximisation Techniques. April 2007. http://arxiv.org/abs/0704.3474. Accessed September 30, 2016.
36. Zoccolella S, Beghi E, Palagano G, et al. Riluzole and amyotrophic lateral sclerosis survival: a population-based study in southern Italy. *Eur J Neurol*. 2007;14(3):262-268. doi:10.1111/j.1468-1331.2006.01575.x.
37. Traynor BJ, Alexander M, Corr B, Frost E, Hardiman O. An outcome study of riluzole in amyotrophic lateral sclerosis. *J Neurol*. 2003;250(4):473-479. doi:10.1007/s00415-003-1026-z.
38. Czaplinski A, Yen AA, Appel SH. Forced vital capacity (FVC) as an indicator of survival and disease progression in an ALS clinic population. *J Neurol Neurosurg Psychiatry*. 2006;77(3):390-392. doi:10.1136/jnnp.2005.072660.

A DEEP LEARNING APPROACH FOR CANCER DETECTION AND RELEVANT GENE IDENTIFICATION

PADIDEH DANAEE*, REZA GHAEINI

School of Electrical Engineering and Computer Science, Oregon State University,
Corvallis, OR 97330, USA
**E-mail: danaeep@oregonstate.edu and ghaeinim@oregonstate.edu*

DAVID A. HENDRIX

School of Electrical Engineering and Computer Science,
Department of Biochemistry and Biophysics, Oregon State University,
Corvallis, OR 97330, USA E-mail: david.hendrix@oregonstate.edu

Cancer detection from gene expression data continues to pose a challenge due to the high dimensionality and complexity of these data. After decades of research there is still uncertainty in the clinical diagnosis of cancer and the identification of tumor-specific markers. Here we present a deep learning approach to cancer detection, and to the identification of genes critical for the diagnosis of breast cancer. First, we used Stacked Denoising Autoencoder (SDAE) to deeply extract functional features from high dimensional gene expression profiles. Next, we evaluated the performance of the extracted representation through supervised classification models to verify the usefulness of the new features in cancer detection. Lastly, we identified a set of highly interactive genes by analyzing the SDAE connectivity matrices. Our results and analysis illustrate that these highly interactive genes could be useful cancer biomarkers for the detection of breast cancer that deserve further studies.

Keywords: Cancer Detection; RNA-seq Expression; Deep Learning; Dimensionality Reduction; Stacked Denoising Autoencoder; Classification.

1. Introduction

The analysis of gene expression data has the potential to lead to significant biological discoveries. Much of the work on the identification of differentially expressed genes has focused on the most significant changes, and may not allow recognition of more subtle patterns in the data.[1–6] Tremendous potential exists for computational methods to analyze these data for the discovery of gene regulatory targets, disease diagnosis and drug development.[7–9] However, the high dimension and noise associated with these data presents a challenge for these tasks. Moreover, the mismatch between the large number of genes and typically small number of samples presents the challenge of a “dimensionality curse”. Multiple algorithms have been used to distinguish normal cells from abnormal cells using gene expression.[10–13] Although there has been a lot of research into cancer detection from gene expression data, there remains a critical need to improve accuracy, and to identify genes that play important roles in cancer.

Machine learning methods for dimensionality reduction and classification of gene expression data have achieved some success, but there are limitations in the interpretation of the most significant signals for classification purposes.[14,15] Recently, there have been efforts to use single-layer, nonlinear dimensionality reduction techniques to classify samples based on gene expression data.[16] In similar studies of computer vision, unsupervised deep learning methods have been successfully applied to extract information from high dimensional image data.[17]

Similarly, one can extract the meaningful part of the expression data by applying such techniques, thereby enabling identification of specific subsets of genes that are useful for biologists and physicians, with the potential to inform therapeutic strategies.

In this work, we used stacked denoising autoencoders (SDAE) to transform high-dimensional, noisy gene expression data to a lower dimensional, meaningful representation.[18] We then used the new representations to classify breast cancer samples from the healthy control samples. We used different machine learning (ML) architectures to observe how the new compact features can be effective for a classification task and allow the evaluation of the performance of different models. Finally, we analyzed the lower-dimensional representations by mapping back to the original data to discover highly relevant genes that could play critical roles and serve as clinical biomarkers for cancer diagnosis. The performance of these methods affirm that SDAEs could be applied to cancer detection in order to improve the classification performance, extract both linear and nonlinear relationships in the data, and perhaps more important, to extract a subset of relevant genes from deep models as a set of potential cancer biomarkers. The identification of these relevant genes deserves further analysis as it potentially can improve methods for cancer diagnosis and treatment.

2. Background

Classification and clustering of gene expression in the form of microarray or RNA-seq data are well studied. There are various approaches for the classification of cancer cells and healthy cells using gene expression profiles and supervised learning models. The self-organizing map (SOM) was used to analyze leukemia cancer cells.[19] A support vector machine (SVM) with a dot product kernel has been applied to the diagnosis of ovarian, leukemia, and colon cancers.[11] SVMs with nonlinear kernels (polynomial and Gaussian) were also used for classification of breast cancer tissues from microarray data.[10]

Unsupervised learning techniques are capable of finding global patterns in gene expression data. Gene clustering represents various groups of similar genes based on similar expression patterns. Hierarchical clustering and maximal margin linear programming are examples of this learning and they have been used to classify colon cancer cells.[20,21] K-nearest neighbors (KNN) unsupervised learning also has been applied to breast cancer data.[12]

Due to the large number of genes, high amount of noise in the gene expression data, and also the complexity of biological networks, there is a need to deeply analyze the raw data and exploit the important subsets of genes. Regarding this matter, other techniques such as principal component analysis (PCA) have been proposed for dimensionality reduction of expression profiles to aid clustering of the relevant genes in a context of expression profiles.[22] PCA uses an orthogonal transformation to map high dimensional data to linearly uncorrelated components.[23] However, PCA reduces the dimensionality of the data linearly and it may not extract some nonlinear relationships of the data.[24] In contrast, other approaches such as kernel PCA (KPCA) may be capable of uncovering these nonlinear relationships.[25]

Similarly, researchers have applied PCA to a set of combined genes of 13 data sets to obtain the linear representation of the gene expression and then apply a autoencoder to capture nonlinear relationships.[26] Recently, a denoising autoencoder has been applied to extract a feature set from breast cancer data.[16] Using a single autoencoder may not extract all the useful representations from the noisy, complex, and high-dimensional expression data. However, by reducing the dimensionality incrementally, the multi-layered architecture of an SDAE may extract meaningful patterns in these data with reduced loss of information.[27]

3. Materials and Methods

We have applied a deep learning approach that extracts the important gene expression relationships using SDAE. After training the SDAE, we selected a layer that has both low-dimension and low validation error compared to other encoder stacks using a validation data set independent of both our training and test set.[28] As a result, we selected an SDAE with four layers of dimensions of 15,000, 10,000, 2,000, and 500. Consequently we used the selected layer as input features to the classification algorithms. The goal of our model is extracting a mapping that possibly decodes the original data as closely as possible without losing significant gene patterns.

We evaluated our approach for feature selection by feeding the SDAE-encoded features to a shallow artificial neural network (ANN)[29] and an SVM model.[30] Furthermore, we applied a similar approach with PCA and KPCA as a comparison.

Lastly, we used the SDAE weights from each layer to extract genes with strongly propagated influence on the reduced-dimension SDAE-encoding. These selected “deeply connected genes” (DCGs) are further tested and analyzed for pathway and Gene Ontology (GO) enrichment. The results from our analysis showed that in fact our approach can reveal a set of biomarkers for the purpose of cancer diagnosis. The details of our method are discussed in the following subsections, and the work-flow of our approach is shown in Fig 1.

3.1. *Gene Expression Data*

For our analysis, we analyzed RNA-seq expression data from The Cancer Genome Atlas (TCGA) database for both tumor and healthy breast samples.[31] These data consist of 1097 breast cancer samples, and 113 healthy samples. To overcome the class imbalance of the data, we used synthetic minority over-sampling technique (SMOTE) to transform data into a more balanced representation for pre-training.[32] We used the `imbalanced-learn` package for this transformation of the training data.[33] Furthermore, we removed all genes that had zero expression across all samples.

3.2. *Dimensionality Reduction Using Stacked Denoising Autoencoder*

An autoencoder (AE) is a feedforward neural network that produces the output layer as close as possible to its input layer using a lower dimensional representation (hidden layer). The autoencoder consists of an encoder and a decoder. The encoder is a nonlinear function, like a sigmoid, applied to an affine mapping of the input layer, which can be expressed as $f_\theta(X) = \sigma(Wx+b)$ with parameters $\theta = \{W, b\}$. The matrix W is of dimensions $d' \times d$ to go from a larger dimension of gene expression data d to a lower dimensional encoding corresponding to d'. The bias vector b is of dimension d'. This input layer encodes the data to generate a

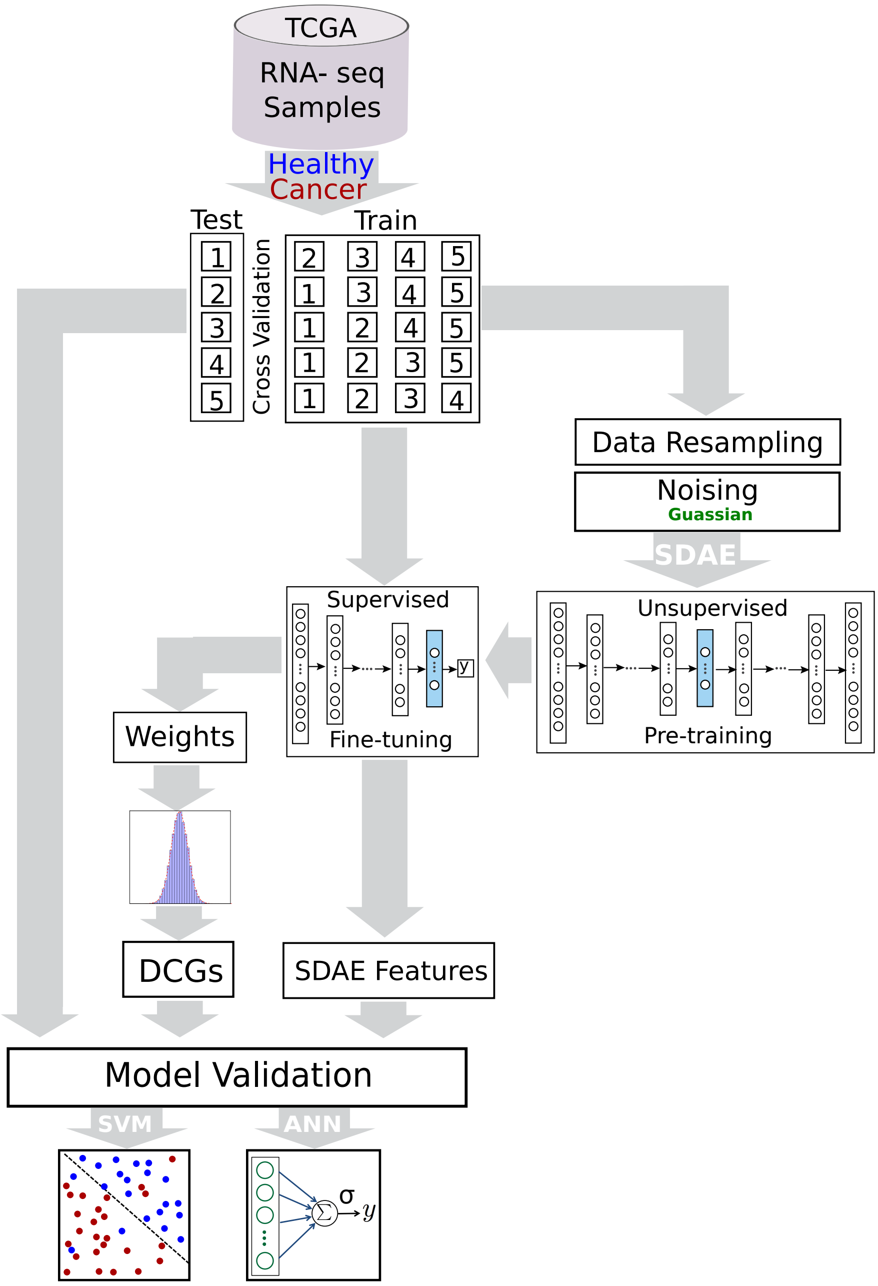


Fig. 1. The pipeline representing the stacked denoising autoencoder (SDAE) model for breast cancer classification and the process of biomarkers extraction.

hidden or latent layer. The decoder takes the hidden representations from the previous layer and decodes the data as closely as possible to the original inputs, and can be expressed as $z = g_{\theta'}(y) = \sigma(W'y + b')$. In our implementation, we imposed tied weights, with $W' = W^T$. We can refer to the weight matrix W and bias b as $\theta = \{W, b\}$ and similarly $\theta' = \{W', b'\}$.

A SDAE can be constructed as a series of AE mappings with parameters $\theta_1, \theta_2, ..., \theta_n$ and the addition of noise to prevent overfitting.[18] In order to get a good representation for each layer, we maximize the information gain between the input layer (modeled as a random variable X from an unknown distribution $q(X)$) and its higher level stochastic representation (random variable Y from a known distribution $p(X|Y;\theta')$). For layer i, we then learned a set of parameters θ_i and θ'_i from a known distribution p where $q(Y|X) = p(Y|X;\theta_i)$ and also $q(X|Y) = p(X|Y;\theta'_i)$ that maximize the mutual information.[18]

This maximization problem corresponds to minimizing the reconstruction error of the input layer using hidden representation. In this construction, the hidden layer contains the compressed information of the data by ignoring useless and noisy features. In fact, the autoencoder extracts a set of new representations which encompass the complex relationships between input variables. The reconstruction error of the input layer using this new representation is non-zero, but can be minimized. In practice, the weights of the model are learned through the stochastic gradient descent (SGD) algorithm.[34,35]

Autoencoders extract both linear and nonlinear relationships inherent in the input data, making them powerful and versatile. The encoder of the SDAE decreases the dimensionality of the gene expression data stack-by-stack, which leads to reduced loss of information compared to reducing the dimension in one step.[27] In contrast, the decoder increases the dimensionality to eventually achieve the full reconstruction of the original input as close as possible. In this procedure, the output of one layer is the input to the next layer. For this implementation, we used the `Keras` library with `Theano` backend running on an Nvidia Tesla K80 GPU.[36] Although it is difficult to estimate the time complexity of the deep architecture of the SDAE, with batch training and highly parallelizable implementation on GPUs, training takes a few minutes and testing of a sample is performed in a few seconds.

It is proven in practice that pre-training the parameters in a deep architecture leads to a better generalization on a specific task of interest.[18] Greedy layer-wise pre-training is an unsupervised approach that helps the model initialize the parameters near a good local minimum and convert the problem to a better form of optimization.[27] Therefore, we considered the pre-training approach as supposed to achieve smoother convergence and higher overall performance in cancer classification. After starting with the initial parameters resulting from the pre-training phase, we used supervised fine-tuning on the full training set to update the parameters.

To avoid overfitting in the learning phase (both pre-training and fine-tuning) of the SDAE, we utilized a dropout regularization factor, which is a method of randomly excluding fractions of hidden units in the training procedure by setting them to zero. This method prevents nodes from co-adapting too much and consequently avoids overfitting.[37] For the same purpose, we provided partially corrupted input values to the SDAE (denoising). The SDAE is robust, and its accuracy does not change upon introducing noise at a low rate. In fact, SDAE with denoising and dropout can find a better representation from the noisy data. Fig 2 shows the SDAE encoded, decoded, and denoised representations on the subset of genes.

3.3. *Differentially Expressed Genes*

We used significantly differentially expressed genes as a comparison to our SDAE features for cancer classification. First, we computed the log fold change comparing the median expression

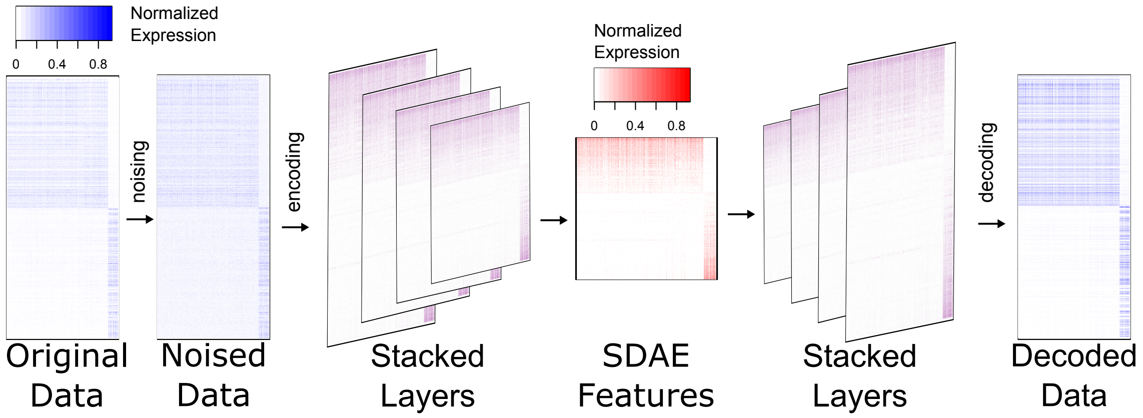


Fig. 2. SDAE representation using the enriched genes in the TCGA breast cancer. In this depiction for illustrative purposes, the top 500 genes with median expression across cancer samples enriched above health samples, and the top 500 genes with reduced median expression across cancer samples is shown.

in cancer tissue samples to that of healthy tissue samples. We then computed a two-tailed p-value using a Gaussian fit, followed by a Benjamini-Hochberg (BH) correction.[38] We identified two sets of differentially expressed genes. The first, DIFFEXP0.05 was the 206 genes, 98 up-regulated and 118 down-regulated, that were significant at an FDR of 0.05. The second set, DIFFEXP500, contains the top 500 most significant differentially expressed genes (the same dimension as the SDAE features) using the same 2-tailed p-values, containing 244 up-regulated and 256 down-regulated genes.

3.4. *Dimensionality Reduction Using Principal Component Analysis*

As a second level of comparison, we extracted features using linear PCA to provide a baseline for the performance of linear dimensionality reduction algorithms for our ML models. The same reduced dimensionality of 500 was used. In addition, we used KPCA with an RBF kernel to extract features that by default are of the same dimension as the number of training input samples. For both PCA and KPCA we used an implementation in the `scikit-learn` package.[39]

4. Results and Discussion

4.1. *Classification Learning*

In order to evaluate the effectiveness of our autoencoder-extracted features, we used two different supervised learning models to classify cancer samples from healthy control samples. First, we considered a single-layer ANN with input nodes directly connected to output layers without any hidden units. If we consider the input units as $X = (x_1, x_2, ..., x_n)$, the output values are calculated as $y = \sigma(\sum_i w_i x_i + b)$. Second, we considered both an SVM with a linear kernel and with a radial basis function kernel (SVM-RBF). We applied 5-fold cross-validation for to exhaustively split the data into train and test sets to estimate the accuracy of each model without overfitting. In each split, the model was trained on 4 partitions and tested on the 5th, ensuring that training and testing are performed on non-overlapping subsets.

5. Comparison of Different Models

To assess the effectiveness of the SDAE features, we compared their performance in classification to differentially expressed genes and to principal components for different machine learning models. The performance of the SDAE features for classification is summarized in Table 1. The best method varies depending on the performance metric, but on these data the SDAE features performed best on three of the five metrics we considered. The highest accuracy was attained using SDAE features applied to SVM-RBF classification. This method also had the highest F-measure. The highest sensitivity was found for SDAE features as well, but using the ANN classification model. KPCA features applied to an SVM-RBF had higher specificity and precision.

Table 1. Comparison of different feature sets using three classification learning models.

Features	Model	Accuracy	Sensitivity	Specificity	Precision	F-measure
SDAE	ANN	96.95	**98.73**	95.29	95.42	0.970
	SVM	98.04	97.21	99.11	99.17	0.981
	SVM-RBF	**98.26**	97.61	99.11	99.17	**0.983**
DIFFEXP500	ANN	63.04	60.56	70.76	84.58	0.704
	SVM	57.83	64.06	46.43	70.42	0.618
	SVM-RBF	77.391	86.69	71.29	67.08	0.755
DIFFEXP0.05	ANN	59.93	59.93	69.95	84.58	0.701
	SVM	68.70	82.73	57.5	65.04	0.637
	SVM-RBF	76.96	87.56	70.48	65.42	0.747
PCA	ANN	96.52	98.38	95.10	95.00	0.965
	SVM	96.30	94.58	98.61	98.75	0.965
	SVM-RBF	89.13	83.31	99.47	99.58	0.906
KPCA	ANN	97.39	96.02	99.10	99.17	0.975
	SVM	97.17	96.38	98.20	98.33	0.973
	SVM-RBF	97.32	89.92	**99.52**	**99.58**	0.943

6. Deep Feature Extraction and Deeply Connected Genes

Going beyond classification, there is potential biological significance in understanding what subsets of genes are involved in the new feature space that makes it an effective set for the cancer detection. Previous work on cancer detection using a single-layer autoencoder has evaluated the importance of each hidden node.[16] Here, we analyzed the importance of genes by considering combined effect of each stack of the deep architecture. To extract these genes, we utilized a strategy of computing the product of the weight matrices for each layer of our SDAE. The result is a $500 \times G$ dimensional matrix W, where G is the number of genes in the expression data, computed for an n-layer SDAE by

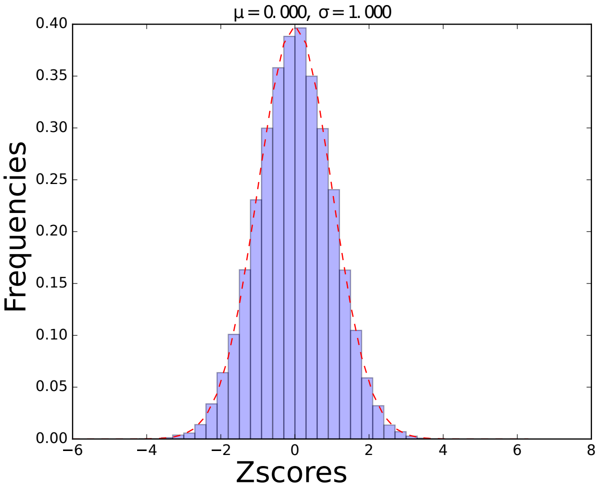


Fig. 3. Histogram of z-Scores from the dot product matrix of the weights connectivity of the SDAE.

$$W = \prod_{i=1}^{n} W_i.$$

Although the weights of each layer of the SDAE are computed with a nonlinear model, the matrix W is a linearization of the compounded effect of each gene on the SDAE features. Genes with the largest weights in W are the most strongly connected to the extracted and highly predictive features, so we called these genes DCGs. We found that the terms of matrix W were strongly normally distributed (Fig 3). We identified the subset of genes with the most statistically significant impact on the encoding by fitting the distribution of these values in W to a normal distribution, computing a p-value using this fit,and applying a BH correction with an FDR of 0.05.

6.1. *Gene Ontology*

We examined the functional enrichment of the DCGs through a GO term and Panther pathway analysis. Table 2 presents the statistically-enriched GO terms under "biological process", and having a Bonferroni-corrected p-value of less than 1e-10. Many of the most significant terms are related to mitosis, suggesting a large number of genes with core functionality that is relevant to cell proliferation. In addition, an analysis of the enrichment of Panther pathways led to a single enriched term, p53 pathway, where we observe 10 genes when 1.34 are expected, giving a p-value of 2.21E-04. P53 is known to be an important tumor-suppressor gene[40–42] , and this finding suggests a role of tumor suppressor function in many of the DCGs.

6.2. *Classification Learning*

Finally, we used the expression of the DCGs as features for the ML models previously mentioned. These genes served as useful features for cancer classification, achieving 94.78% accuracy (Table 3). Although these features performed a few percentage points below that of the

Table 2. Enriched GO terms associated with DCGs in breast cancer data from TCGA.

GO biological process	Total	Observed	Expected	Enrichment	P-value
cell cycle process (GO:0022402)	1079	100	16.46	6.07	1.12E-45
cell cycle (GO:0007049)	1311	108	20	5.4	3.28E-45
mitotic cell cycle process (GO:1903047)	741	85	11.31	7.52	1.06E-44
mitotic cell cycle (GO:0000278)	760	85	11.6	7.33	7.33E-44
nuclear division (GO:0000280)	470	63	7.17	8.78	1.52E-35
organelle fission (GO:0048285)	492	64	7.51	8.53	1.99E-35
mitotic nuclear division (GO:0007067)	357	56	5.45	10.28	1.34E-34
cell division (GO:0051301)	477	58	7.28	7.97	3.72E-30
chromosome segregation (GO:0007059)	274	46	4.18	11	5.46E-29
sister chromatid segregation (GO:0000819)	176	36	2.69	13.41	7.62E-25
nuclear chromosome segregation (GO:0098813)	230	38	3.51	10.83	3.97E-23
mitotic cell cycle phase transition (GO:0044772)	249	35	3.8	9.21	8.10E-19
mitotic prometaphase (GO:0000236)	99	25	1.51	16.55	1.56E-18
cell cycle phase transition (GO:0044770)	255	35	3.89	9	1.72E-18
regulation of cell cycle (GO:0051726)	943	62	14.39	4.31	2.48E-18
chromosome organization (GO:0051276)	984	63	15.01	4.2	4.15E-18
DNA metabolic process (GO:0006259)	768	52	11.72	4.44	2.67E-15
organelle organization (GO:0006996)	3133	112	47.8	2.34	4.27E-15
mitotic cell cycle phase (GO:0098763)	211	29	3.22	9.01	7.67E-15
cell cycle phase (GO:0022403)	211	29	3.22	9.01	7.67E-15
biological phase (GO:0044848)	215	29	3.28	8.84	1.25E-14
sister chromatid cohesion (GO:0007062)	113	22	1.72	12.76	1.18E-13
cellular resp. to DNA damage stimu. (GO:0006974)	719	48	10.97	4.38	1.27E-13
regulation of cell cycle process (GO:0010564)	557	42	8.5	4.94	2.53E-13
mitotic sister chromatid segregation (GO:0000070)	90	20	1.37	14.56	3.01E-13
cell cycle checkpoint (GO:0000075)	196	25	2.99	8.36	1.09E-11
M phase (GO:0000279)	173	23	2.64	8.71	6.55E-11
mitotic M phase (GO:0000087)	173	23	2.64	8.71	6.55E-11
regulation of mitotic cell cycle (GO:0007346)	461	35	7.03	4.98	1.10E-10
single-organism process (GO:0044699)	12451	253	189.98	1.33	4.67E-10
DNA replication (GO:0006260)	213	24	3.25	7.38	5.74E-10
anaphase (GO:0051322)	154	21	2.35	8.94	6.13E-10
mitotic anaphase (GO:0000090)	154	21	2.35	8.94	6.13E-10
cellular component organization (GO:0016043)	5133	139	78.32	1.77	7.74E-10

SDAE features, they still have advantage of being more readily interpreted. Future work is needed to improve the extraction of DCGs to enhance their utility as features for classification.

Table 3. Cancer classification results using deeply connected genes (DCGs).

Features	Model	Accuracy	Sensitivity	Specificity	Precision	F-measure
DCGs	ANN	91.74	98.13	87.15	85.83	0.913
	SVM	91.74	88.80	97.50	97.25	0.927
	SVM-RBF	94.78	93.04	97.5	97.20	0.951

6.3. *Conclusion*

In conclusion, we have used a deep architecture, SDAE, for the extraction of meaningful features from gene expression data that enable the classification of cancer cells. We were able to use the weights of this model to extract genes that were also useful for cancer prediction, and have potential as biomarkers or therapeutic targets.

One limitation of deep learning approaches is the requirement for large data sets, which may not be available for cancer tissues. We expect that as more gene expression data becomes available, this model will improve in performance and reveal more useful patterns. Accordingly, deep learning models are highly scalable to large input data.

Future work is needed to analyze different types of cancer to identify cancer-specific biomarkers. In addition, there is potential to identify cross-cancer biomarkers through the analysis of aggregated heterogeneous cancer data.

References

1. E. Kettunen, S. Anttila, J. K. Seppänen, A. Karjalainen, H. Edgren, I. Lindström, R. Salovaara, A.-M. Nissén, J. Salo, K. Mattson *et al.*, *Cancer genetics and cytogenetics* **149**, 98 (2004).
2. C. G. A. R. Network *et al.*, *Nature* **499**, 43 (2013).
3. J. Xu, J. A. Stolk, X. Zhang, S. J. Silva, R. L. Houghton, M. Matsumura, T. S. Vedvick, K. B. Leslie, R. Badaro and S. G. Reed, *Cancer research* **60**, 1677 (2000).
4. H. Li, B. Yu, J. Li, L. Su, M. Yan, J. Zhang, C. Li, Z. Zhu and B. Liu, *PloS one* **10**, p. e0125013 (2015).
5. T. Zhou, Y. Du and T. Wei, *Biophysics Reports* **1**, 106 (2015).
6. J. S. Myers, A. K. von Lersner, C. J. Robbins and Q.-X. A. Sang, *PloS one* **10**, p. e0145322 (2015).
7. M. Maienschein-Cline, J. Zhou, K. P. White, R. Sciammas and A. R. Dinner, *Bioinformatics* **28**, 206 (2012).
8. E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang *et al.*, *Nature genetics* **37**, 710 (2005).
9. K. Shabana, K. A. Nazeer, M. Pradhan and M. Palakal, *BMC bioinformatics* **16**, p. 1 (2015).
10. S. Reddy, K. T. Reddy, V. V. Kumari and K. V. Varma, *International Journal of Computer Science and Information Technologies* **5**, 5901 (2014).
11. T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer and D. Haussler, *Bioinformatics* **16**, 906 (2000).
12. S. A. Medjahed, T. A. Saadi and A. Benyettou, *International Journal of Computer Applications* **62** (2013).
13. A. C. Tan and D. Gilbert (2003).
14. J. A. Cruz and D. S. Wishart, *Cancer informatics* **2** (2006).
15. K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis and D. I. Fotiadis, *Computational and structural biotechnology journal* **13**, 8 (2015).
16. C. C. e. a. Tan J, Ung M, Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders (2015).
17. H. Lee, R. Grosse, R. Ranganath and A. Y. Ng, Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.
18. P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio and P.-A. Manzagol, *The Journal of Machine Learning Research* **11**, 3371 (2010).

19. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri *et al.*, *science* **286**, 531 (1999).
20. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack and A. J. Levine, *Proceedings of the National Academy of Sciences* **96**, 6745 (1999).
21. J. Li, H. Liu, S.-K. Ng and L. Wong, *Bioinformatics* **19**, ii93 (2003).
22. K. Y. Yeung and W. L. Ruzzo, *Bioinformatics* **17**, 763 (2001).
23. S. Wold, K. Esbensen and P. Geladi, *Chemometrics and intelligent laboratory systems* **2**, 37 (1987).
24. A. Gupta, H. Wang and M. Ganapathiraju, Learning structure in gene expression data using deep architectures, with an application to gene clustering, in *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, 2015.
25. B. Schölkopf, A. Smola and K.-R. Müller, Kernel principal component analysis, in *International Conference on Artificial Neural Networks*, 1997.
26. R. Fakoor, F. Ladhak, A. Nazi and M. Huber, Using deep learning to enhance cancer diagnosis and classification, in *Proceedings of the ICML Workshop on the Role of Machine Learning in Transforming Healthcare. Atlanta, Georgia: JMLR: W&CP*, 2013.
27. Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle *et al.*, *Advances in neural information processing systems* **19**, p. 153 (2007).
28. G. E. Hinton and R. R. Salakhutdinov, *Science* **313**, 504 (2006).
29. S.-C. Wang, Artificial neural network, in *Interdisciplinary Computing in Java Programming*, (Springer, 2003) pp. 81–100.
30. C. Cortes and V. Vapnik, *Machine learning* **20**, 273 (1995).
31. J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network *et al.*, *Nature genetics* **45**, 1113 (2013).
32. N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, *Journal of artificial intelligence research* **16**, 321 (2002).
33. G. Lemaître, F. Nogueira and C. K. Aridas, *CoRR* **abs/1609.06570** (2016).
34. D. Saad, *Online Learning* .
35. O. Bousquet and L. Bottou, The tradeoffs of large scale learning, in *Advances in neural information processing systems*, 2008.
36. F. Chollet, Keras `https://github.com/fchollet/keras`, (2015).
37. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, *The Journal of Machine Learning Research* **15**, 1929 (2014).
38. Y. Benjamini and Y. Hochberg, *Journal of the royal statistical society. Series B (Methodological)* , 289 (1995).
39. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *Journal of Machine Learning Research* **12**, 2825 (2011).
40. G. Matlashewski, P. Lamb, D. Pim, J. Peacock, L. Crawford and S. Benchimol, *The EMBO journal* **3**, p. 3257 (1984).
41. M. Isobe, B. Emanuel, D. Givol, M. Oren and C. M. Croce (1986).
42. S. E. Kern, K. W. Kinzler, A. Bruskin, D. Jarosz, P. Friedman, C. Prives and B. Vogelstein, *Science* **252**, 1708 (1991).

DEVELOPMENT AND PERFORMANCE OF TEXT-MINING ALGORITHMS TO EXTRACT SOCIOECONOMIC STATUS FROM DE-IDENTIFIED ELECTRONIC HEALTH RECORDS

BRITTANY M. HOLLISTER

Vanderbilt Genetics Institute, Vanderbilt University, 519 Light Hall, 2215 Garland Ave. South
Nashville, TN, 37232, USA
Email: Brittany.M.Hollister@Vanderbilt.edu

NICOLE A. RESTREPO

Institute for Computational Biology, Department of Epidemiology and Biostatistics, Case Western Reserve University,
Wolstein Research Building, 2103 Cornell Road, Cleveland, OH 44106, USA
Email: nrestrepo@case.edu

ERIC FARBER-EGER

Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, 2525 West End
Avenue, Suite 600, Nashville, TN 37203, UA
Email: eric.h.farber-eger@vanderbilt.edu

DANA C. CRAWFORD

Institute for Computational Biology, Department of Epidemiology and Biostatistics, Case Western Reserve University,
Wolstein Research Building, 2103 Cornell Road, Suite 2527, Cleveland, OH 44106, USA
Email: dana.crawford@case.edu

MELINDA C. ALDRICH[†]

Department of Thoracic Surgery and Division of Epidemiology, Vanderbilt University Medical Center, 1313 21st
Avenue South, 609 Oxford House, Nashville, TN 37232, USA
Email: melinda.aldrich@vanderbilt.edu

AMY NON[†]

Department of Anthropology, University of California, San Diego, 9500 Gilman Drive #0532
La Jolla, CA 92093, USA
Email: alnon@ucsd.edu

[†] Co-Senior authors

Socioeconomic status (SES) is a fundamental contributor to health, and a key factor underlying racial disparities in disease. However, SES data are rarely included in genetic studies due in part to the difficultly of collecting these data when studies were not originally designed for that purpose. The emergence of large clinic-based biobanks linked to electronic health records (EHRs) provides research access to large patient populations with longitudinal phenotype data captured in structured fields as billing codes, procedure codes, and prescriptions. SES data however, are often not explicitly recorded in structured fields, but rather recorded in the free text of clinical notes and communications. The content and completeness of these data vary widely by practitioner. To enable gene-environment studies that consider SES as an exposure, we sought to extract SES variables from racial/ethnic minority adult patients (n=9,977) in BioVU, the Vanderbilt University Medical Center biorepository linked to de-identified EHRs. We developed several measures of SES using information available within the de-identified EHR, including broad categories of occupation, education, insurance status, and homelessness. Two hundred patients were randomly selected for manual review to develop a set of seven algorithms for extracting SES information from de-identified EHRs. The algorithms consist of 15 categories of information, with 830 unique search terms. SES data extracted from manual review of 50 randomly selected records were compared to data produced by the algorithm, resulting in positive predictive values of 80.0% (education), 85.4% (occupation), 87.5% (unemployment), 63.6% (retirement), 23.1% (uninsured), 81.8% (Medicaid), and 33.3% (homelessness), suggesting some categories of SES data are easier to extract in this EHR than others. The SES data extraction approach developed here will enable future EHR-based genetic studies to integrate SES information into statistical analyses. Ultimately, incorporation of measures of SES into genetic studies will help elucidate the impact of the social environment on disease risk and outcomes.

1. Introduction

1.1. *Socioeconomic status and health*

Socioeconomic status (SES) is a major determinant of variation in health outcomes worldwide[1]. SES is typically defined as a combination of income or wealth, educational achievement, and occupation[2,3] and be can assessed at the individual, household, or neighborhood level. Health outcomes within the United States, ranging from cancer to hypertension, vary by socioeconomic levels, regardless of how they are measured[4]. Multiple measures of SES have been previously associated with health outcomes, including income[5], years of education[6,7], occupational prestige[2,8,9], insurance coverage[10], and homelessness[11].

SES likely affects health through various pathways including access to healthcare services, knowledge of health behaviors, exposure to environmental stressors and hazards, limited financial resources, and social support[2]. The relationship between SES and health is also highly entangled with race/ethnicity, as SES may covary with race and contribute in part to the existence of racial disparities in health[4,12]. Though these pathways are difficult to distinguish and could affect different populations to varying degrees, it is important to consider SES as a representation of these potential pathways in studies of human health.

Despite the overwhelming evidence that SES affects health outcomes, SES measures are often not included in genetic studies of disease. Neglect of SES data may be due to a lack of available SES information in existing cohorts, as well as the additional time and resources it takes to collect SES data in new studies. Despite these difficulties, it is vital to include measurements of SES in genetic association studies of racial disparities in health. In addition to the possible confounding that may occur due to the association of race/ethnicity with both SES and health[13], SES has the potential to modify the effect of genetic variants on health outcomes[14]. Therefore, the etiology of disease is likely to be misunderstood without the inclusion of SES data in association studies. Although prior genetic association studies have found some gene variants that may explain a small portion of racial disparities in disease prevalence and risk[15], SES factors are likely to play an even larger role in racial health disparities[6,7].

1.2. *SES data within electronic health records*

The use of electronic health records (EHRs) for research purposes is becoming increasingly prevalent. With the announcement of the Precision Medicine Initiative and its goal of recruiting one million participants with biological and EHR data, the research use of EHRs is anticipated to increase[16]. EHRs provide an attractive resource for biomedical researchers for many reasons, including their rich phenotypic and longitudinal data, as well as the lower cost of participant recruitment versus a traditional prospective cohort study. Additionally, clinical biobanks that contain biological samples linked to EHRs are becoming an invaluable resource for conducting genetic epidemiology studies. Despite the potential for EHRs in research settings, these clinical data repositories currently have noted deficits in the availability and completeness of important social and environmental data[17], including SES, that are known to contribute independently to health status and could modify genetic effects[18].

Recognizing the importance of formally and consistently capturing social and behavioral measures in the EHR, the Institute of Medicine (IOM) recently recommended SES measures, specifically educational attainment, financial resource strain, and neighborhood median household income be included in the EHR[19]. The committee also recommended that a plan be developed by the NIH to expand the research use of EHRs to include social and behavioral data[19]. Adoption of these recommendations will take time, and may not be universal across medical centers; therefore, there is a need to develop approaches and methods to access existing unstructured SES data within the EHR for research purposes. SES data are almost entirely found within the free text clinical notes from providers and the clinical communications between providers. Currently, there are no published algorithms available to extract SES data from EHRs. In this study, we developed an approach for extracting available SES information from the free text of a de-identified EHR. These algorithms will facilitate the immediate extraction of key SES information from clinical biobanks for incorporation into future biomedical research.

1.3. *BioVU*

BioVU is a DNA biobank of the Vanderbilt University Medical Center (VUMC) linked to de-identified EHRs. DNA samples are extracted from discarded blood samples drawn for routine clinical care[20]. DNA samples are linked to the Synthetic Derivative (SD), the de-identified version of the VUMC EHR, by a unique study ID. Medical records within the SD are scrubbed of all HIPAA identifiers such as names, locations, zip codes, and social security numbers. Dates within each SD record are shifted to prevent re-identification of the records. Date shifting is consistent within a single patient’s record. As previously described[21], data from BioVU are de-identified in accordance with provisions of Title 45, Code of Federal Regulations, part 46 (45 CFT 46); consequently, this study is considered non-human subjects research by the Vanderbilt University Institutional Review Board.

2. Methods

2.1. *Population*

The study population included all racial/ethnic minority patients >18 years old participating in BioVU as of 2011[22]. The EHRs used for the development of the algorithms were updated in 2015 to include current information. Race/ethnicity is administratively reported in BioVU and strongly correlated with genetic ancestry[23,24]. The majority (81%) of patients in the dataset are black individuals with an average age of 50 years (Table 1). The mean number of clinic visits within a patient's EHR record is 40.45 visits, and the mean number of days between patients' first and last visit within the EHR is 2,340 days (Table 1).

Table 1. Vanderbilt BioVU racial/ethnic minority population characteristics

Characteristic	n= 9,977
Sex	
Male	3,568 (36%)
Female	6,409 (64%)
Race/ethnicity	
Black	8,078 (81%)
Hispanic	1,049 (10.5%)
Asian	850 (8.5%)
Age (mean, years ± SD)	49.8 ± 18.1
Number of clinic visits (mean ± SD)	40.5 ± 55.0
Number of days between visits (mean ± SD)	2,340 ± 1,793.1

2.2. *Development of algorithms*

We sought to develop algorithms to extract SES data from structured and unstructured data in the de-identified EHRs. We developed seven algorithms for the extraction of SES information including education level, occupation, unemployment, retirement, insurance status, Medicaid status, and homelessness (Table 2). The initial development of the SES algorithms began with a manual review of both structured and unstructured data within the de-identified EHR of 200 randomly selected patients within this minority population dataset to identify the following: 1) the categories of SES information most frequently mentioned, 2) where in the EHR this information is noted, and 3) the semantic language used by clinical providers for socioeconomic information (Figure 1). The manual review revealed that the SES data were found exclusively within the unstructured free text of the clinical notes, social history, and clinical communications of this EHR. It was also noted that the most frequently mentioned semantic categories were employment, education, insurance status, and homelessness, and thus these categories were chosen for extraction. Semantic tags for each category were selected if they appeared more than once within the 200 development records.

2.2.1. *Employment*

Employment information was extracted using three different algorithms designed to capture data on occupation, unemployment, and retirement. The occupation algorithm extracts the occupation

Table 2. Variables extracted by socioeconomic status (SES) algorithms applied to de-identified electronic health records

Semantic category	Format of algorithm output
Occupational prestige	0-100
Unemployment	Ever/never
Retirement	Ever/never
Education	-Never attended -Less than high school -High school graduate/GED -Associate's degree -Bachelor's degree -Master's degree -Professional degree -Doctoral degree
Uninsured	Ever/never
Medicaid	Ever/ never
Homelessness	Ever/never

mentioned in a patient's record and translates it to an occupational prestige score (scale of 0-100). This score represents how well-respected an occupation is within a society (i.e., subjective socioeconomic position). Occupational prestige scores were determined from a National Opinion Research Center (NORC) survey where respondents were asked to rank occupations according to their prestige[25]. The occupation tags utilized for the occupation algorithm were adopted from the most recent NORC report. The algorithm's occupation tags were shortened to 678 occupations from the original NORC list of 860 occupations given that some of the occupations were highly specific with repetitive occupational prestige scores. As an example, "teacher, elementary school" and "teacher, secondary school" were collapsed to "teacher."

We next used the occupation algorithm to search the unstructured data of the original 200 patients for the initial occupation tags. This search identified a large number of false positives, where the algorithm tagged occupation-related words that were not indicative of the patient's occupation. In order to filter these false positives, additional prefix language such as "works as," "is a/n," "employed" was added to a subset of occupations to filter out non-relevant terms, which greatly improved the algorithm. Unemployment data were extracted using semantic tags for unemployment (e.g., "unemployed," "does not work," "hasn't worked since"). The unemployment algorithm was then tested on the unstructured data from the 200 records used for development, and a high number of false positives were returned. These false positives were often in reference to medications. Therefore the tags "if this does not work" and "if that does not work" were excluded to filter false positives. Unemployment was classified as ever/never (Table 2). Retirement was also extracted from the EHR using the tag "retired" and classified as ever/never (Table 2).

2.2.2. *Education*

The education algorithm was designed to assign education level to a patient based on the highest education achieved and recorded in the EHR. Education levels were assigned to each relevant tag word or phrase found in the unstructured text of the EHR (Table 2). Sixty-two semantic tags were utilized and the highest level of education was determined for each patient. These tags were exclusive to an assigned education level. For example, the high school degree category of education level included tags such as “high school graduate” and “completed 12th grade,” while the bachelor’s degree category included terms such as “BS degree” and “completed college.” The levels of education were based on U.S. census definitions with one modification such that all grade levels below high school graduate were collapsed into a “less than high school” category. We searched through the unstructured text of the 200 records used for development to determine if further filtering or modification was needed. Fifteen additional tags were used to filter false positive results related to types of medical education (e.g. “diet education,” “dialysis education”) and Vanderbilt Medical School students (e.g., “medical student,” “pharmacy student,” “student nurse”).

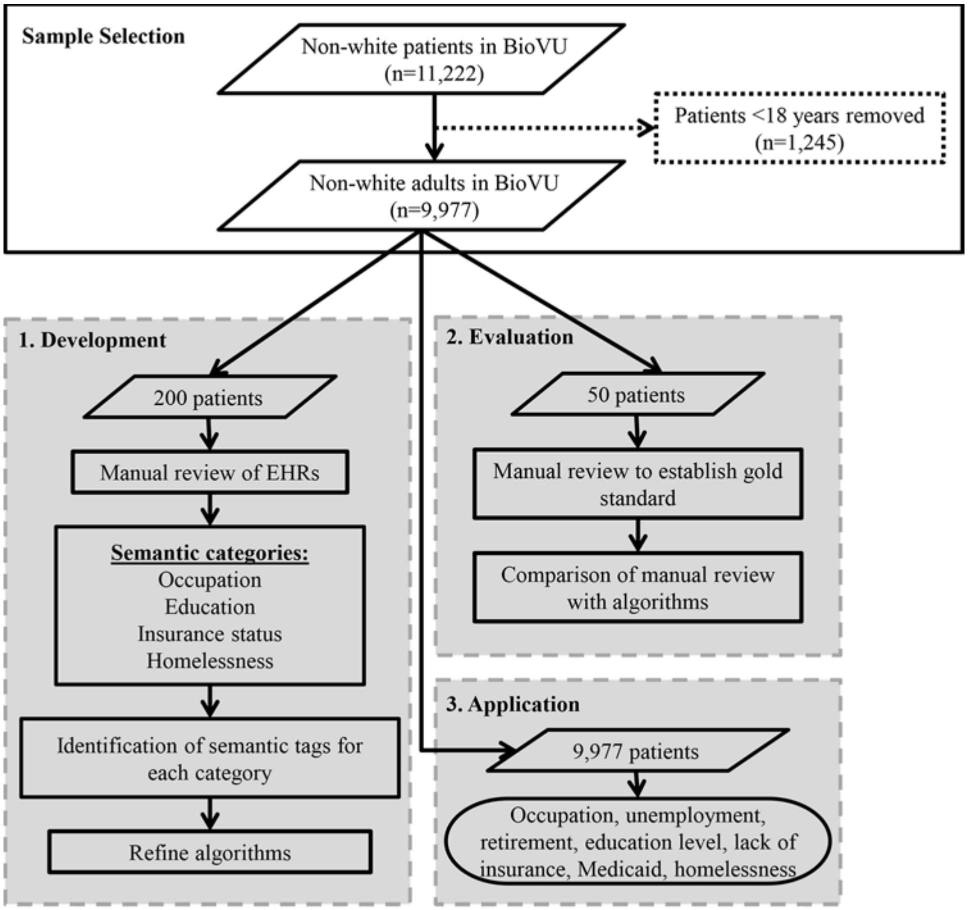

Sample Selection
Non-white patients in BioVU (n=11,222)
Patients <18 years removed (n=1,245)
Non-white adults in BioVU (n=9,977)
1. Development
200 patients
Manual review of EHRs
Semantic categories:
Occupation
Education
Insurance status
Homelessness
Identification of semantic tags for each category
Refine algorithms
2. Evaluation
50 patients
Manual review to establish gold standard
Comparison of manual review with algorithms
3. Application
9,977 patients
Occupation, unemployment, retirement, education level, lack of insurance, Medicaid, homelessness

Figure 1. Overview of the development process for the SES algorithms

2.2.3. *Insurance status*

The extraction process for insurance status required two algorithms. The first algorithm was used to determine if there was any time point in the EHR when the patient did not have insurance based on the presence of five semantic tags (Table 2). These tags included “no insurance” and “does not have insurance” and excluded some language that was used in a standard discharge letter and therefore appeared frequently in the EHR. A second insurance algorithm extracted Medicaid information using specific phrases or keywords such as “Medicaid” and “TennCare” and was classified as ever/never in order to determine if a patient was ever on Medicaid in their EHR (Table 2).

2.2.4. *Homelessness*

Homelessness information was extracted using the tags “homeless” and “shelter” among the 200 development EHRs. After this search, several false positives were returned relating to patients who worked or volunteered at homeless shelters. Therefore, exclusion tags were added such as “volunteer at homeless shelter,” “works at homeless shelter,” “works with homeless,” and “animal shelter.” Homelessness was classified as ever/never (Table 2).

2.3. *Evaluation of algorithm performance*

To evaluate the performance of these SES algorithms, results were compared to findings from a manual review of 50 randomly selected patients. These 50 individuals were selected using random sampling without replacement. Two independent reviewers manually reviewed the clinical record of each patient and any discrepancies were resolved by discussion between the two reviewers. Comparison of results from the two independent reviewers was quantified using percent positive agreement, percent negative agreement, and kappa statistics for each of the seven categories and subcategories: education level, occupation, unemployment, retirement, uninsured, Medicaid, and homelessness. The manual review of 50 records was then compared to the algorithm results for each of the seven categories and subcategories. Sensitivity, specificity, and positive predictive value were estimated. The chi-square statistic was used to determine if the algorithms performed differently in different populations.

3. Results

3.1. *Population characteristics*

Among the total study population (n=9,977), we were able to extract at least one type of SES information from 8,282 (83.0%) individuals. We extracted education information for 3,780 individuals and occupation information for 7,296 individuals (Table 3). For the remaining

Table 3. Percent of records within the study population with algorithm-identified SES characteristics

Characteristics	Race			
	Black (n=8,078)	Hispanic (n=1,049)	Asian (n=850)	Total (n=9,977)
% with occupation	76.0	57.1	65.4	73.1
% unemployed	21.4	13.0	13.4	19.8
% retired	19.8	4.9	11.2	17.5
% with education	39.1	28.7	37.9	37.9
% uninsured	19.5	15.6	11.5	18.4
% on Medicaid	20.5	13.9	7.9	18.7
% homeless	3.7	1.3	1.0	3.2

categories, we were able to determine if an individual was unemployed, retired, uninsured, on Medicaid, or homeless at any point in his or her record. Of the total population for which we were able to extract SES data (n=8,282), 1,978 individuals were unemployed, 1,742 individuals were retired, 1,839 individuals were uninsured, 1,865 were on Medicaid, and 318 were homeless at least one time in their EHR (Table 3). For each of the seven categories, the algorithms returned SES information for a higher percentage of black patients than Hispanic or Asian patients (p<0.00001).

The five most frequently extracted occupations among those having occupation information (n=7,296) were manager, nurse, Army, manufacturer, and restaurant employee. Within the population with education information (n=3,780), the vast majority of individuals had a high school degree (n=2,066), followed by individuals without a high school degree (n=492), and individuals with a bachelor's degree (n=446).

3.2. *Algorithm Performance*

Prior to evaluating algorithm performance, the manual review results from the randomly selected records of 50 patients were compared between the two reviewers and any conflicts were resolved. The percent positive agreement between reviewers ranged from 98.0% to 100.0% and the percent negative agreement ranged from 94.7% to 100.0%. The Kappa statistic between reviewers ranged from 0.94 to 1.0.

Table 4. Comparison of manual review with algorithm results for each SES algorithm in a subset of randomly selected individuals (n=50)

Semantic Category	Records with SES information (%)	Sensitivity (%)	Specificity (%)	PPV (%)
Education level	48.0	66.7	84.5	80.0
Occupation	80.0	87.5	40.0	85.4
Unemployment	40.0	70.0	93.3	87.5
Retirement	14.0	100.0	90.7	63.6
Uninsured	8.00	75.0	78.3	23.1
Medicaid	18.0	100.0	95.1	81.8
Homelessness	2.00	100.0	95.9	33.3

Once all reviewer discrepancies were resolved, the manual review results were used as the gold standard and compared to the algorithm results. All the algorithms, with the exception of occupation, had very high specificity levels >78%. The lower specificity for occupation (40%) is due to six of the ten individuals who did not have occupation information (as identified by manual review) but were identified as having occupation information by the algorithm. All the algorithms also had high sensitivity levels (above 70%), with the exception of education level (66.7%) (Table 4). The lower sensitivity for education is driven by eight individuals who have an education level that was identified by manual review but not by the algorithm. The lower sensitivity for unemployment is due to the six individuals who were identified as unemployed by manual review but not by the algorithm. PPV values across the algorithms ranged from 23.1%-87.5%. The lower PPV for the retirement algorithm (63.6%) is due to the four individuals identified as retired by the algorithm but not retired by the manual review. (Table 4). The low PPV for the uninsured algorithm (23.1%) is due to the ten individuals who were identified as uninsured by the algorithm, but not by manual review. The low PPV for homelessness (33.3%) was a result of the fact that the manual review only identified one patient with homelessness in their record, whereas the algorithm misidentified two others.

3.2.1. *Missing data*

Of the total population (n=9,977), the algorithm was not able to extract any SES information for 1,695 individuals (17.0%). Of this group, there were 1,193 blacks, 309 Hispanics, and 193 Asians. Missing SES data were more common among Hispanic and Asian individuals, than among black individuals ($p<0.001$). The Hispanic and Asian populations represent 10.5% and 8.5% of the total dataset, respectively; however, these groups represent 18.2% and 11.4%, respectively, of the individuals with missing SES data. Males represent 35.8% of the study population and 28.0% of those without extracted SES data. The mean age for the total population is 49.9 years, and the mean age for the group without extracted SES information is 46.7 years.

4. Conclusion

Socioeconomic status is considered a fundamental cause of disease, because it affects so many proximate risk factors and disease outcomes[26]. SES has been consistently associated with health outcomes such as mortality, cancer, and cardiovascular disease[27,28]. Despite these consistent associations, SES data are typically not included in genetic studies of health outcomes. For studies that utilize biobanks, the lack of SES data is likely related to the difficulty in accessing these data within the EHR, where they are not usually recorded in structured fields. The algorithms described in this study are the first to extract these important data from EHRs for research purposes.

The SES algorithms described here focus on the extraction of data related to four semantic categories: occupation, education, insurance status, and homelessness. The occupation algorithms extracted and classified data as occupational prestige, unemployment (ever/never), and retirement (ever/never). The occupational prestige algorithm had a strong sensitivity and PPV; however it had a low specificity of 40%, reflective of the difficulty in filtering the occupation information. Although we took steps to remove false positives, it was difficult to completely eliminate all false positives without removing a large amount of accurate data. Our unemployment and retirement algorithms had high sensitivity (70% and 100%) and specificity (93.3% and 90.7%). While the unemployment algorithm had a high PPV, the retirement algorithm had a low PPV. Both unemployment and retirement were classified as ever/never because the EHR only captures a snapshot of time when the patient visits the clinic. It was not possible to accurately capture the length of time for unemployment or retirement as the patient's visits to the clinic may not reflect the length of time he or she was unemployed or retired. The sensitivity of the unemployment algorithm was affected by the varying language used to describe unemployment, which was identified in manual review but not consistently recognized by the algorithm ("does not work outside the home", "used to work in a restaurant"). The quality of the retirement algorithm was affected by false positives related to the identification of words related to retirement that were used in a context outside of the patient's retirement from an occupation.

The education algorithm identified the highest level of education that a patient achieved over the course of their EHR. This algorithm had a high specificity and PPV, but a low sensitivity. The low sensitivity was due to the inability of the algorithm to detect variations in education level compared with the manual review. The variation in language used by clinical providers made it difficult to include every mention of education while still maintaining some level of precision. For example, some of the Vanderbilt Medical School students were excluded ("medical student," "pharmacy student") because of the frequent mention of these terms in the EHR related to patient care, rather than education level. The reviewers were able to infer education level based on occupation and context clues as well as identify the medical school students, while the algorithm was not able to so. The algorithm that identified patients who were uninsured at some point in his or her record as well as the homelessness algorithm each had high sensitivity and specificity, but low PPV. Uninsured patients are the smallest portion of patients within VUMC, making up only 4.7% of the patient population in 2015[29]. The low PPV of these algorithms may be due to a low prevalence of uninsured patients and homeless individuals within the VUMC patient population. Within our randomly selected minority patient population used for evaluation, only four individuals were uninsured and one was homeless. These categories had the lowest prevalence within our evaluation dataset. The Medicaid algorithm was the highest performing algorithm, with a high sensitivity, specificity, and PPV.

The major challenges in utilizing EHR data in a research setting include missing data and the inconsistencies in the recording of SES data by clinical providers. While the majority of individuals within the study population had some SES information, a notable percentage of individuals did not have any SES information within their records (17.0%). The missing SES data could be a result of the lack of recording of information by the provider, either due to SES factors not being discussed in conversation with the patient, a low number of visits in the patient's EHR, or the willingness of the patient to provide SES information. Additionally, when variables are missing within a patient's record, we cannot distinguish whether it is due to negative data or just missing data. For example, if a patient does not have an occupation listed, we cannot assume that they are unemployed because it may have not been discussed with the provider or recorded by the provider. The higher level of missing data observed for Hispanic and Asian individuals in this dataset could be a reflection of the fact that the algorithms are optimized for the largest racial/ethnic population within the dataset (i.e., black patients).

The inconsistencies in the recording of the SES data are typical for social and environmental exposure data contained within free clinical text[17]. In the development of these algorithms, we noted that providers, in general, do not follow patterns when recording SES data within their notes in the EHR. The lack of consistent language and the numerous variations used to describe the SES information made extracting this information challenging. Furthermore, algorithms could also be limited by the accuracy of the selected filters and tags, rather than the information available within the EHR. While the aim of the algorithms was to include all possible semantic tags, there is a possibility that some information was missed by the algorithms or that information was captured inaccurately due to the limitations of the filtering process.

In addition to these general limitations, the algorithms developed here have specific limitations regarding portability. Even within the same dataset, we have noted a difference in tag retrieval for the SES categories queried across the three major racial/ethnic groups. Additional studies are required to improve the algorithms' performances and retrieval of semantic tags in multiple populations as well as within different study sites. Indeed, some of the tags developed here (such as "TennCare" in reference to Medicaid) are specific to Tennessee and will require modification to ensure portability regardless of the state in which the algorithms are deployed. Furthermore, these algorithms were created in a de-identified EHR, which required the development of a free text algorithm for insurance status, as the structured insurance information is considered identifying information. An identified EHR may have this insurance information within the structured text. However, the other categories of SES information are likely to only be found within the free text of an identified EHR.

Despite the many challenges faced with the extraction of SES data from the EHR, these algorithms were able to successfully extract a large amount of data not previously accessible for research purposes. The sensitivities, specificities, and PPVs for the algorithms were high considering the limitations of the SES data within the current EHR. Overall, these algorithms represent a first important step in incorporating SES data from EHRs into precision medicine research, as envisioned by the Institute of Medicine and others.

5. Resources

Semantic tag and filter lists for each algorithm can be found on the Vanderbilt University Medical Center TREAT Lung Cancer Research Program website (https://medschool.vanderbilt.edu/treat-lung-cancer-program/) and the Institute for Computational Biology website (http://www.icompbio.net/?page_id=1654). The code which was used to run the algorithms is available in GitHub.

6. Acknowledgements

This work was supported in part by NIH grants U01 HG004798 and its ARRA supplements (DCC) and 1K07CA172294 (MCA). BMH was supported by the NIH/NIGMS Genetics Predoctoral Research Training Program 5T32GM080178-07. The dataset(s) used for the analyses described were obtained from Vanderbilt University Medical Center's BioVU which is supported by institutional funding and by the Vanderbilt CTSA grant funded by the National Center for Research Resources, Grant UL1 RR024975-01, which is now at the National Center for Advancing Translational Sciences, Grant 2 UL1 TR000445-06.

References

1. *Poverty: Assessing the Distribution of Health Risks by Socioeconomic Position at National and Local Levels.* (2004).
2. T. Seeman *et al.*, *Social Science & Medicine* 66, 72-87 (2008).
3. P. Braveman *et al.*, *Public Health Reports* 129 Suppl 2, 19-31 (2014).
4. National Center for Health Statistics, *Health, United States, 2011: With Special Feature on Socioeconomic Status and Health* (2012).

5. V. Carrieri *et al.*, *Health Econ*, (2016).
6. A. L. Non *et al.*, *American Journal of Public Health* 102, 1559-1565 (2012).
7. M. C. Aldrich *et al.*, *American Journal of Public Health* 103, e73-80 (2013).
8. R. Hauser *et al.*, *Sociological Methodology* 27, 177-298 (1997).
9. K. Fujishiro *et al.*, *Social Science & Medicine* 71, 2100-2107 (2010).
10. in *Kaiser Commission on Medicaid and the Uninsured* T. H. J. K. F. Foundation, Ed. (Washington, D.C., 2012).
11. D. S. Morrison, *International Journal of Epidemiology* 38, 877-883 (2009).
12. National Center for Health Statistics *Health, United States, 2015: With Special Feature on Racial and Ethnic Health Disparities* (2016).
13. T. J. VanderWeele *et al.*, *Epidemiology* 25, 473-484 (2014).
14. S. Cakmak *et al.*, *Journal of Environmental Management* 177, 1-8 (2016).
15. J. S. Kaufman *et al.*, *American Journal of Epidemiology* 181, 464-472 (2015).
16. F. S. Collins *et al.*, *The New England Journal of Medicine* 372, 793-795 (2015).
17. I. S. Kohane, *Nature Reviews. Genetics* 12, 417-428 (2011).
18. J. Basson *et al.*, *American journal of hypertension* 27, 431-444 (2014).
19. IOM (Institute of Medicine), *Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2* (2014).
20. D. M. Roden *et al.*, *Clinical Pharmacology and Therapeutics* 84, 362-369 (2008).
21. J. Pulley *et al.*, *Clinical and Translational Science* 3, 42-48 (2010).
22. D. C. Crawford *et al.*, *Human Heredity* 79, 137-146 (2015).
23. J. B. Hall *et al.*, *PloS one* 9, e99161 (2014).
24. L. Dumitrescu *et al.*, *Genetics in Medicine : official journal of the American College of Medical Genetics* 12, 648-650 (2010).
25. NORC, *Measuring Occupational Presitge on the 2012 General Social Survey* (2014).
26. B. G. Link *et al.*, *J Health Soc Behav* Spec No, 80-94 (1995).
27. T. N. Bethea *et al.*, *Ethnicity & Disease* 26, 157-164 (2016).
28. A. Rawshani *et al.*, *JAMA Internal Medicine*, (2016).
29. "2015 Financial Report " (Vanderbilt University, Nashville, TN. , 2015).

GENOME-WIDE INTERACTION WITH SELECTED TYPE 2 DIABETES LOCI REVEALS NOVEL LOCI FOR TYPE 2 DIABETES IN AFRICAN AMERICANS

JACOB M. KEATON[1,2,3], JACKLYN N. HELLWEGE[2,3], MAGGIE C. Y. NG[2,3], NICHOLETTE D. PALMER[2,3,4,5], JAMES S. PANKOW[6], MYRIAM FORNAGE[7], JAMES G. WILSON[8], ADOLFO CORREA[8], LAURA J. RASMUSSEN-TORVIK[9], JEROME I. ROTTER[10], YII-DER I. CHEN[10], KENT D. TAYLOR[10], STEPHEN S. RICH[11], LYNNE E. WAGENKNECHT[5,12], BARRY I. FREEDMAN[3,5,13], DONALD W. BOWDEN[2,3,4]

[1]Molecular Genetics and Genomics Program, Wake Forest School of Medicine, Medical Center Blvd, Winston-Salem, NC, 27157, US

[2]Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Medical Center Blvd, Winston-Salem, NC, 27157, US

[3]Center for Diabetes Research, Wake Forest School of Medicine, Wake Forest School of Medicine, Medical Center Blvd, Winston-Salem, NC, 27157, US

[4]Department of Biochemistry, Wake Forest School of Medicine, Medical Center Blvd, Winston-Salem, NC, 27157, US

[5]Center for Public Health Genomics, Wake Forest School of Medicine, Medical Center Blvd, Winston-Salem, NC, 27157, US

[6]Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, MN, 55455, US

[7]Institute of Molecular Medicine and Human Genetics Center, University of Texas Health Science Center at Houston, 7000 Fannin St #1200, Houston, TX, 77030, US

[8]University of Mississippi Medical Center, 2500 N State St, Jackson, MS, 39216, US

[9]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, 303 E Chicago Ave, Chicago, IL, 60611, US

[10]Institute for Translational Genomics and Population Sciences, Los Angeles BioMedical Research Institute, Harbor-UCLA Medical Center, 1000 W Carson St, Torrance, CA, 90502, US

[11]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, 22904, US

[12]Division of Public Health Sciences, Wake Forest School of Medicine, Medical Center Blvd, Winston-Salem, NC, 27157, US

[13]Department of Internal Medicine - Section on Nephrology, Wake Forest School of Medicine, Medical Center Blvd, Winston-Salem, NC, 27157, US

Type 2 diabetes (T2D) is the result of metabolic defects in insulin secretion and insulin sensitivity, yet most T2D loci identified to date influence insulin secretion. We hypothesized that T2D loci, particularly those affecting insulin sensitivity, can be identified through interaction with known T2D loci implicated in insulin secretion. To test this hypothesis, single nucleotide polymorphisms (SNPs) nominally associated with acute insulin response to glucose (AIR_g), a dynamic measure of first-phase insulin secretion, and previously associated with T2D in genome-wide association studies (GWAS) were identified in African Americans from the Insulin Resistance Atherosclerosis Family Study (IRASFS; n=492 subjects). These SNPs were tested for interaction, individually and jointly as a genetic risk score (GRS), using GWAS data from five cohorts (ARIC, CARDIA, JHS, MESA, WFSM; n=2,725 cases, 4,167 controls) with T2D as the outcome. In single variant analyses, suggestively significant ($P_{\text{interaction}} < 5\times10^{-6}$) interactions were observed at several loci including *DGKB* (rs978989), *CDK18* (rs12126276), *CXCL12* (rs7921850), *HCN1* (rs6895191), *FAM98A* (rs1900780), and *MGMT* (rs568530). Notable beta-cell GRS interactions included two SNPs at the *DGKB* locus (rs6976381; rs6962498). These data support the hypothesis that additional genetic factors contributing to T2D risk can be identified by interactions with insulin secretion loci.

1. Introduction

Although common variants examined in genome-wide association studies (GWAS) have identified ~80 loci associated with T2D risk, these variants explain only about 15% of T2D heritability[1,2]. A portion of the missing heritability may be explained by epistasis, which occurs when a genetic risk factor is modified by other factors in an individual's genetic background[3]. Epistasis, or gene-gene interaction, analyses may facilitate the detection of novel loci when non-additive effects exist, but may also provide novel insights illuminating biological mechanisms underlying complex diseases such as T2D[4].

T2D is characterized by impaired insulin secretion arising from pancreatic beta-cell dysfunction and insulin resistance in skeletal muscle, hepatic, and other peripheral tissues, leading to decreased plasma glucose uptake. However, documented T2D loci primarily map to genes influencing insulin secretion or other aspects of beta-cell biology[1]. Given the underlying bimodal pathophysiology, T2D may be a particularly well-suited disease model for hypothesis-driven investigation of epistatic interactions. Genetic insults to both insulin secretion and insulin sensitivity may jointly increase an individual's T2D risk in a non-additive manner. Considering the higher prevalence rate of T2D, insulin resistance, and obesity, African Americans are optimal for the study of genetic interactions that contribute to T2D risk.

In an effort to identify interactions contributing to T2D and to discover novel insulin sensitivity loci, we hypothesized that T2D risk loci, particularly those affecting insulin sensitivity, could be identified by interaction analyses with known T2D loci implicated in insulin secretion. In cross-sectional meta-analyses of five T2D studies (ARIC, CARDIA, JHS, MESA, and WFSM), we tested whether 5 SNPs from known T2D loci implicated in insulin secretion, or a genetic risk score summarizing these SNPs, modified genome-wide SNP associations with T2D risk.

2. Research Design and Methods

2.1 *Subjects*

Two sources of data were analyzed in this study. Primary inferences of association with insulin secretion were derived from African American participants (n=492 individuals from 42 families) in the Insulin Resistance Atherosclerosis Family Study (IRASFS), a metabolically well-characterized cohort[5]. Glucose homeostasis traits were measured by the frequently sampled intravenous glucose tolerance test (FSIGT)[5]. Briefly, a 50% glucose solution (0.3g/kg) and regular human insulin (0.03units/kg) were injected intravenously at 0 and 20 minutes, respectively. Blood was collected at −5, 2, 4, 8, 19, 22, 30, 40, 50, 70, 100, and 180 minutes for measurement of plasma glucose and insulin. AIR_g was calculated as the increase in insulin at 2–8 minutes above the basal (fasting) insulin level after the bolus glucose injection at 0-1 minute. Insulin sensitivity (S_I) was calculated by mathematical modeling using the MINMOD program (version 3.0 [1994])[6]. Disposition index (DI) was calculated as the product of S_I and AIR_g.

Inferences of genome-wide epistatic interaction with insulin secretion loci for T2D susceptibility were derived from African American participants from the Atherosclerosis Risk in Communities Study (ARIC; n = 955 T2D cases, 414 controls), Coronary Artery Risk Development in Young Adults (CARDIA; n = 94 T2D cases, 654 controls), Jackson Heart Study (JHS; n = 333 T2D cases, 1,450 controls), Multi-Ethnic Study of Atherosclerosis (MESA; n = 411 T2D cases, 793 controls), and the Wake Forest School of Medicine (WFSM; n = 932 T2D cases, 856 controls) cohorts for a total of 2,725 T2D cases and 4,167 controls[7–12]. T2D was diagnosed according to the American Diabetes Association criteria with at least one of the following: fasting glucose ≥126 mg/dL, 2-h oral glucose tolerance test glucose ≥200 mg/dL, random glucose ≥200 mg/dL, use of oral hypoglycemic agents and/or insulin, or physician diagnosed diabetes. Subjects diagnosed before 25 years of age were excluded. Normal glucose tolerance was defined as fasting glucose <100 mg/dL and 2-h oral glucose tolerance test glucose <140 mg/dL (if available) without reported use of diabetes medications. Control subjects <25 years of age were excluded.

IRB approval was obtained at all sites and all participants provided written informed consent. Descriptions of the T2D study cohorts are summarized in the Supplementary Methods.

2.2 *Genotyping, imputation, and quality control*

For the IRASFS samples, genotyping and quality control were performed at the Wake Forest Center for Genomics and Personalized Medicine Research using the Illumina Infinium HumanExome BeadChip v1.0 as previously described[13]. Briefly, the exome chip contained 247,870 variants (92% protein coding). In addition, the chip included 64 SNPs associated with T2D from previous GWAS in Europeans, many of which have been implicated in insulin secretion (exome chip design: http://genome.sph.umich.edu/wiki/Exome_Chip_Design). Sample and autosomal SNP call rates were ≥99%, and SNPs with poor cluster separation (<0.35) were excluded. Mendelian errors were identified using PedCheck[14] and resolved by removing conflicting genotypes. Hardy–Weinberg Equilibrium (HWE) was assessed in unrelated samples (n = 39) using PLINK (http://pngu.mgh.harvard.edu/purcell/plink)[15] to reduce biases introduced by familial allele frequencies. All variants were in accordance with HWE ($P > 1 \times 10^{-5}$).

The T2D study samples were genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0. For the ARIC, CARDIA, JHS, and MESA cohorts, genotyping and quality control were completed by the National Heart, Lung, and Blood Institute's (NHLBI's) Candidate Gene Association Resource (CARe) at the Broad Institute[16]. Genotyping for the WFSM study was performed at the Center for Inherited Disease Research (CIDR). For all T2D studies, imputation was performed using MACH with the function –mle (version 1.0.16, http://www.sph.umich.edu/csg/abecasis/MaCH/) to obtain missing genotypes and replace genotypes inconsistent with reference haplotypes as previously described[17]. SNPs with call rate ≥ 95% and minor allele frequency (MAF) ≥ 1% that passed study-specific quality control were used for imputation[16,18]. A 1:1 HapMap II (NCBI Build 36) CEU:YRI (European:African) consensus haplotype was used as reference. A total of 2,713,329 to 2,907,086 autosomal SNPs from each GWAS with call rate ≥95%, MAF ≥ 1%, and Hardy-Weinberg P-value ≥ 0.0001 for genotyped SNPs and MAF ≥ 1% and RSQ ≥ 0.5 for imputed SNPs were included in subsequent data analyses.

2.3 *Principal component analysis*

For IRASFS, admixture was estimated using principal components (PCs) from 39 ancestry informative markers (AIMs) and including HapMap CEU and YRI samples for comparison[19]. Only PC1 correlated with HapMap populations, and was thus used as a covariate in all analyses.

For the T2D studies, PCs were computed for each study using high-quality SNPs as previously described[13,16–18,20]. The first PC was highly correlated ($r^2 > 0.87$) with global African-European ancestry, as measured by ANCESTRYMAP[21], STRUCTURE[22], or FRAPPE[23]. The African American T2D study samples had an average of 80% African ancestry. By analyzing unrelated samples from all studies using SMARTPCA[20], only the first PC appeared to account for substantial genetic variation (data not shown), whereas the subsequent PCs may reflect sampling noise and/or relatedness in samples[21]. The first PC (PC1) was used as a covariate in all analyses to adjust for population substructure.

2.4 *Analysis of association with measures of glucose homeostasis in IRASFS*

To approximate a normal distribution, trait values were transformed by square root (AIR_g, DI) or natural logarithm plus a constant (S_I). Measured genotype association analyses of exome chip variants with AIR_g, S_I, and DI were performed under an additive model using the variance components method implemented in Sequential Oligogenic Linkage Analysis Routines (SOLAR)[24] with adjustment for age, gender, body mass index (BMI), and PC1.

2.5 *Genetic risk score construction*

We further explored our interaction approach by constructing genetic risk scores (GRS), both weighted and unweighted, summarizing the effects of SNPs associated with both T2D and insulin secretion (T2D-IS SNPs). The T2D-IS GRS was created using the T2D risk alleles for T2D-IS SNPs defined from the literature (Table 1). The unweighted risk score was calculated by summation of the number of risk alleles for each individual across all selected SNPs. The weighted T2D-IS GRS was calculated as the sum of risk alleles at each locus multiplied by the natural log of their T2D odds ratio (OR) defined from the literature[2,25–28]. Missing genotypes for a given SNP were imputed as the average number of risk alleles across all samples. The association of each GRS with both AIR_g and DI, a combinatorial measure of first-phase insulin secretion and insulin sensitivity, were evaluated in IRASFS using the variance components method implemented in SOLAR[24] adjusted for age, gender, and ancestry proportions.

Table 1. Characteristics and single-SNP AIR_g association results for T2D-IS SNPs in published GWAS and IRASFS

T2D-IS SNP	Chr	Position*	Gene	Published GWAS				IRASFS AIR_g			
				T2D Risk Allele	Other Allele	T2D OR†	PMID‡	RAF§	Beta	SE‖	P
rs7593730	2	161171454	RBMS1	T	C	1.11	20418489	0.39	-1.38	0.86	0.086
rs864745	7	28180556	JAZF1	T	C	1.10	18372903	0.72	-1.52	0.91	0.096
rs5215	11	17408630	KCNJ11	C	T	1.08	24509480	0.15	-2.60	1.18	0.033
rs1552224	11	72433098	ARAP1	A	C	1.14	20581827	0.06	-3.05	1.69	0.077
rs7119	15	77777632	HMG20A	C	T	1.24	22885922	0.52	-1.50	0.81	0.059

*NCBI build 37. †Reported odds ratio. ‡PubMed ID. §Risk allele frequency. ||Standard error.

2.6 *Analysis of interaction for T2D risk in the African American T2D case-control studies*

A logistic regression test for additive allelic interaction adjusted for age, gender, and PC1 was used for all interaction analyses with T2D as the outcome. Additional models included adjustment for BMI, and individuals with missing values were excluded (n = 110). In each study, genome-wide interaction tests were performed in PLINK between each SNP in the genome with each candidate SNP (i.e. insulin secretion SNP) and GRS (i.e. insulin secretion risk score). An example PLINK command is provided in the Supplementary Methods. Interaction results with extreme values (absolute β or SE > 10), primarily due to low cell counts, were excluded. Across interaction analyses with all SNPs and risk scores, the number of SNPs excluded as outliers ranged from 0 to 17,000. Interaction results were combined by fixed-effect inverse variance weighting for each candidate SNP or GRS in METAL (http://www.sph.umich.edu/csg/abecasis/metal/). Each meta-analysis contained results for 486,148 to 2,965,304 SNPs.

3. Results

3.1 *Candidate beta-cell function SNP selection*

The characteristics of IRASFS subjects are shown in Supplementary Table 1. Samples included 492 African Americans with mean age 41.2 years and mean BMI 29.1 kg/m^2. Average African ancestry proportion was 0.75. FSIGT was performed for all subjects without T2D (n = 492) to assess measures including insulin secretion (AIR_g), insulin sensitivity index (S_I), and disposition index (DI).

We identified 5 SNPs (Table 1) from established T2D risk loci from published GWAS[25–28] in which the T2D risk alleles were trending towards association ($P < 0.10$) with AIR_g in IRAS-FS (T2D-IS SNPs). Selected SNPs were identical to the published T2D GWAS index SNPs with the exception of rs7119 (*HMG20A*), which is in strong linkage disequilibrium with the GWAS index SNP rs7178572 in the current study ($r^2 \geq 0.73$ in all cohorts) and is suggestively associated with T2D ($P = 5.24\text{x}10^{-7}$) in individuals from Southeast Asia[29].

3.2 *Interaction analysis*

The selected SNPs were examined for genome-wide first order multiplicative interactions with 1) individual insulin secretion SNPs and 2) risk scores summarizing these insulin secretion SNPs. To maximize power, these analyses were performed in five studies (ARIC, CARDIA, JHS, MESA, and WFSM) including 2,725 T2D cases and 4,167 non-diabetic controls and results were meta-analyzed. Representative meta-analysis q-q plots are provided in Supplementary Figures 1 and 2. A flowchart summarizing experimental workflow is provided in Supplementary Figure 3.

The characteristics of T2D case (n = 2,725) and control subjects (n = 4,167) for each study cohort are shown in Supplementary Table 2. Mean age at examination ranged from 38.2 (CARDIA) to 67.6 (MESA) years. Mean age at diagnosis for T2D cases ranged from 35.0 (CARDIA) to 54.6 (MESA) years. In all cohorts except WFSM, BMI was >3 kg/m^2 higher in cases compared to controls.

3.3 *T2D-IS SNP interactions*

Five T2D-IS SNPs were tested for genome-wide interactions for T2D risk in the ARIC, CARDIA, JHS, MESA, and WFSM cohorts. Individual T2D-IS SNP results were meta-analyzed across cohorts. While no interactions were observed at a genome-wide significance level, a total of 21 SNP-pairs demonstrated suggestive evidence of interaction ($P_{\text{interaction}} < 5\text{x}10^{-6}$; Table 2). The most significant T2D-IS SNP interaction observed was between rs7119 at the *HMG20A* locus (T2D-IS SNP) and rs6487610 (interacting SNP; $P_{\text{interaction}} = 3.83\text{x}10^{-7}$). This interacting SNP is located in an intron of *SMCO2*, which encodes single-pass membrane protein with coiled-coil domains 2. Top interactions with T2D-IS SNPs overall were robust against BMI adjustment (Table 2), with similar p-values. Other notable interacting SNPs included rs978989 (*DGKB*), rs12126276 (*CDK18*), rs7921850 (*CXCL12*), rs6895191 (*HCN1*), rs1900780 (*FAM98A*), and rs568530 (*MGMT*).

Table 2. Top meta-analyzed interactions with T2D-IS SNPs regressed on T2D risk in ARIC, CARDIA, JHS, MESA, and WFSM

T2D-IS SNP (Gene)	Intxn SNP* (Gene)	Chr	Position†	MAF‡	β_{intxn}§	P_{intxn}§	P_{het}‖	$\beta_{\text{intxn\_adj\_bmi}}$¶	$P_{\text{intxn\_adj\_bmi}}$¶
rs5215 (*KCNJ11*)	rs3024370 (*F13A1*)	6	6250967	0.48	-0.52	3.01E-06	0.71	-0.56	2.32E-06
rs5215 (*KCNJ11*)	rs7842913 (*FUT10*)	8	33089041	0.07	-2.77	4.58E-06	1.00	-2.75	4.57E-06
rs7119 (*HMG20A*)	rs12121207 (*ATG4C*)	1	63232384	0.44	-0.29	2.68E-06	0.20	-0.28	1.43E-05
rs7119 (*HMG20A*)	rs1900780 (*FAM98A*/*MYADML*)	2	33901094	0.33	0.36	3.46E-06	0.76	0.37	6.92E-06
rs7119 (*HMG20A*)	rs978989 (*DGKB*)	7	14954759	0.27	0.33	2.72E-06	0.23	0.33	4.27E-06
rs7119 (*HMG20A*)	rs6487610 (*SMCO2*)	12	27628742	0.38	0.32	3.83E-07	0.42	0.32	8.45E-07
rs7119 (*HMG20A*)	rs7965793 (*ANKS1B*)	12	100175468	0.31	0.44	1.05E-06	0.76	0.47	7.74E-07
rs7119 (*HMG20A*)	rs1496811 (Intergenic)	18	38952563	0.49	0.27	4.95E-06	0.98	0.27	1.24E-05
rs7119 (*HMG20A*)	rs4812424 (Intergenic)	20	38654372	0.35	-0.47	4.68E-07	0.14	-0.46	1.51E-06
rs7119 (*HMG20A*)	rs6105151 (*ESF1*)	20	13691752	0.34	0.30	2.08E-06	0.42	0.32	7.23E-07
rs7593730 (*RBMS1*)	rs6895191 (*HCN1*)	5	45877674	0.28	0.32	2.80E-06	0.39	0.32	6.91E-06

rs7593730 (*RBMS1*)	rs4705321 (*SH3TC2*/*ABLIM3*)	5	148508860	0.31	0.30	4.13E-06	0.58	0.28	2.91E-05
rs7593730 (*RBMS1*)	rs16872382 (*ZFPM2*)	8	106108691	0.03	-0.97	7.34E-07	0.85	-0.99	8.49E-07
rs7593730 (*RBMS1*)	rs12865410 (Intergenic)	13	104785227	0.35	-0.30	9.69E-07	0.46	-0.32	6.44E-07
rs7593730 (*RBMS1*)	rs12863474 (Intergenic)	13	104784409	0.37	0.33	1.29E-06	0.89	0.36	4.48E-07
rs864745 (*JAZF1*)	rs12126276 (*CDK18*)	1	205494508	0.18	-0.92	1.31E-06	0.68	-0.92	2.98E-06
rs864745 (*JAZF1*)	rs12343907 (*GLT6D1*)	9	138498904	0.35	-0.34	1.44E-06	0.87	-0.34	2.04E-06
rs864745 (*JAZF1*)	rs7921850 (*CXCL12*)	10	44704401	0.37	-0.33	2.52E-06	0.56	-0.31	1.37E-05
rs864745 (*JAZF1*)	rs568530 (*MGMT*)	10	131018864	0.41	0.32	3.27E-06	0.30	0.32	1.03E-05
rs864745 (*JAZF1*)	rs16973790 (*WRD72*/*UNC13C*)	15	54188148	0.15	0.55	3.13E-06	0.27	0.51	3.09E-05
rs864745 (*JAZF1*)	rs12483006 (*SLC37A1*)	21	43953851	0.07	-0.66	1.95E-06	0.58	-0.64	8.17E-06

*SNP interacting with selected T2D-IS SNP. †NCBI build 37. ‡Minor allele frequency. §Meta-analyzed effect size and p-value from interaction models adjusted for age, gender, and PC1. ||Heterogeneity p-values across studies from interaction models adjusted for age, gender, and PC1. ¶ Meta-analyzed effect size and p-value from interaction models adjusted for age, gender, PC1, and BMI.

3.4 *GRS validation and interaction analysis*

Each GRS was tested for association with AIR_g and DI under an additive model using the variance components method with adjustment for age, gender, and PC1 in IRASFS (Supplementary Table 3). The weighted T2D-IS GRS was not associated with AIR_g; it was associated with DI with or without BMI adjustment (P = 4.43x10^{-2} and 4.51x10^{-2}, respectively). Since the weighted risk score was associated with measures of glucose homeostasis, analysis of this risk score was emphasized in the tests for genome-wide interaction in the ARIC, CARDIA, JHS, MESA, and WFSM cohorts.

Meta-analyzed estimates of genome-wide interactions with the weighted T2D-IS GRS are presented in Table 3. No interactions met conventional GWAS thresholds for significance. However, eight interactions with the weighted T2D-IS GRS reached a suggestive level of significance ($P_{\text{interaction}}$ < 5x10^{-6}; Table 3). The most significant T2D-IS GRS interaction was with rs12434405 (Table 3, $P_{\text{interaction}}$ = 9.60x10^{-7}). This is an intronic SNP in the gene *CEP128*, which encodes centrosomal protein 128kDa. Further, the T2D-IS GRS interaction analysis identified two SNPs at the *DGKB* locus, rs6976381 and rs6962498 ($r^2 \geq 0.75$ in all cohorts). This locus was identified in single variant interaction analyses with T2D-IS SNP rs7119 (*HMG20A*), though through a different interacting SNP (rs978989). Two SNPs at the *FAM98A* locus, rs6543772 and rs11687252, were also identified in this analysis. This locus was implicated in single variant analyses with T2D-IS SNP rs7119 (*HMG20A*) through the interacting SNP rs1900780. Top interactions with the T2D-IS GRS were also robust against BMI adjustment.

Table 3. Top meta-analyzed interactions with weighted T2D-IS GRS regressed on T2D risk in ARIC, CARDIA, JHS, MESA, and WFSM

Intxn SNP* (Gene)	Chr	Position†	MAF‡	β_{intxn}§	P_{intxn}§	P_{het}||	$\beta_{\text{intxn\_adj\_bmi}}$¶	$P_{\text{intxn\_adj\_bmi}}$¶
rs6543722 (*FAM98A*)	2	33832523	0.39	-1.20	2.82E-06	0.79	-1.22	3.52E-06
rs11687252 (*FAM98A*)	2	33834496	0.38	-1.17	3.27E-06	0.68	-1.19	3.70E-06
rs6851672 (*DKK2*)	4	107907908	0.03	3.70	4.79E-06	0.82	3.63	9.62E-06
rs6976381 (*DGKB*)	7	15048814	0.18	-1.67	1.21E-06	0.73	-1.66	2.18E-06

rs6962498 (*DGKB*)	7	15050305	0.14	-1.77	3.71E-06	0.54	-1.77	6.65E-06
rs17082105 (*PCDH9*)	13	67685156	0.18	1.45	3.46E-06	0.86	1.51	2.65E-06
rs12434405 (*CEP128*)	14	81044614	0.12	-1.90	9.60E-07	0.12	-1.87	2.49E-06
rs16951940 (Intergenic)	16	80021664	0.03	3.40	2.29E-06	0.84	3.43	4.58E-06

*SNP interacting with the weighted T2D-IS GRS. †NCBI build 37. ‡Minor allele frequency. §Meta-analyzed effect size and p-value from interaction models adjusted for age, gender, and PC1. ||Heterogeneity p-values across studies from interaction models adjusted for age, gender, and PC1. ¶ Meta-analyzed effect size and p-value from interaction models adjusted for age, gender, PC1, and BMI.

4. Discussion

Meta-analyses of five African American T2D studies did not reveal genome-wide statistically significant ($P_{\text{interaction}} < 5\times10^{-8}$) first-order interactions with insulin secretion SNPs or composite risk scores. However, the observed interactions ($P_{\text{interaction}} < 5\times10^{-6}$) suggest that a candidate insulin secretion SNP/GRS interaction approach is a valid method for identifying insulin sensitivity and T2D risk loci. For example, analyses with the T2D-IS SNP rs864745 (*JAZF1*) revealed an interaction with rs7921850, an intergenic SNP downstream of the *CXCL12* gene encoding chemokine (C-X-C motif) ligand 12 (also known as stromal cell-derived factor 1). CXCL12 is an adipocyte-derived chemotactic factor that recruits macrophages and is required for the establishment of obesity-induced adipose tissue inflammation and systemic insulin resistance in mice[30].

Several genes related to pancreatic beta-cell function were also identified; suggesting interactions are not limited to insulin resistance as in our initial hypothesis. Evaluations of the T2D-IS SNP rs7119 (*HMG20A*) and the T2D-IS GRS identified interactions with rs978989 and rs6976381, respectively, intergenic SNPs downstream of the *DGKB* gene. Variants at *DGKB* have been associated with T2D, fasting glucose, and pancreatic islet beta-cell function as measured by HOMA-B[27,31]. Variants near *DGKB* disrupt islet-specific enhancer activity[32]. Several other variants detected in our analyses show interactions with similar biological relationships to insulin secretion and T2D.

Interestingly, we observed interactions discrete for individual loci. For example, analyses with rs864745 (*JAZF1*), a locus involved in transcriptional repression, showed an interaction with rs568530, an intergenic SNP upstream of *MGMT*, which encodes O-6-Methylguanine-DNA Methyltransferase. These observations may reflect different, input-dependent physiological characteristics of interaction results, and may lead to mechanistic insights about the underlying causes of T2D and defects in glucose homeostasis in expanded analyses.

Although results varied widely between interaction analyses, interactions with two loci, *DGKB* and *FAM98*, were replicated in multiple analyses. Functional characteristics of *FAM98* related to T2D and glucose homeostasis pathophysiology are not evident in the current literature.

Previous GWAS have largely ignored epistatic contributions to T2D risk due to the heavy multiple testing burden and computational challenges of exhaustive analytical approaches, and when they have considered this contribution, results have not been striking. For example, a recent genome-wide scan for two-locus interactions in the Wellcome Trust Case Control Consortium T2D GWAS data did not reveal any significant epistatic signals at a Bonferroni-corrected p-value threshold of 2.14×10^{-11} after adjusting for the main effects of the most strongly associated T2D locus, *TCF7L2*[33]. Further, Herold et al. estimated that analysis of all pairwise interactions among 550,000 SNPs in 1,200 samples on a 3 GHz computer would require a running time of 120 days[34]. The interaction analysis presented here overcomes the issue of a heavy multiple testing burden by using a candidate SNP approach. A recent study by Becker et al. demonstrated that a multiple test correction of 0.4m, where m is the number of SNP pairs tested, is sufficiently conservative for large-scale allelic interaction tests[35]. Further, Babron et al. show that a correction for the effective number of SNP pairs is equally sufficient[36]. Li et al. previously demonstrated that the effective number of SNPs for an imputed dataset is $\sim 10^6$. These findings suggest that a significance threshold of 1×10^{-8} is appropriate for this study.

We did not detect interactions even at the conventional GWAS threshold of 5×10^{-8} in the current study. In part, this likely reflects the challenge of inherently reduced power of interaction models due to the low frequency of compound genotypes[37]. Computational resources required for this study were equivalent to the requirements for running 12 GWAS (5 candidate insulin secretion SNPs plus a GRS, with and without BMI adjustment). This is a significant reduction compared to exhaustive approaches examining genome-wide interactions with all available SNP pairs.

In summary, our findings demonstrate that genome-wide interaction studies with selected insulin secretion variants is a powerful approach for the detection of T2D risk, insulin secretion, and insulin sensitivity loci. The use of a high-quality measure of first-phase insulin secretion, AIR_g, to identify candidate interaction SNPs yielded compelling associations. These results justify an expansion of the current study and further investigation of putative insulin sensitivity loci, namely *CXCL12*.

Acknowledgements

The authors would like to acknowledge the contributions of the involved research institutions, study investigators, field staff, and study participants of ARIC, CARDIA, JHS, MESA, and WFSM.

Genotyping services for the WFSM study were provided by CIDR. CIDR is fully funded through a federal contract from the National Institutes of Health (NIH) to The Johns Hopkins University (Contract HHSC268200782096C). The work at Wake Forest was supported by NIH grants K99-DK-081350 (N.D.P.), R01-DK-066358 (D.W.B.), R01-DK-053591 (D.W.B.), R01-HL-56266 (B.I.F.), and R01-DK-070941 (B.I.F.), and in part by the General Clinical Research Center of the WFSM Grant M01-RR-07122. This work was also supported by the NHLBI.

The following four parent studies have contributed parent study data, ancillary study data, and DNA samples through the Massachusetts Institute of Technology-Broad Institute (N01-HC-65226) to create this genotype/phenotype database for wide dissemination to the biomedical research community: ARIC, CARDIA, JHS, and MESA.

The Atherosclerosis Risk in Communities (ARIC) Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human

Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. The authors thank the staff and participants of the ARIC study for their important contributions. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research. The authors thank the staff and participants of the ARIC study for their important contributions.

The Coronary Artery Risk Development in Young Adults (CARDIA) Study is conducted and supported by the National Heart, Lung, and Blood Institute in collaboration with the University of Alabama at Birmingham (HHSN268201300025C & HHSN268201300026C), Northwestern University (HHSN268201300027C), University of Minnesota (HHSN268201300028C), Kaiser Foundation Research Institute (HHSN268201300029C), and Johns Hopkins University School of Medicine (HHSN268200900041C). CARDIA is also partially supported by the Intramural Research Program of the National Institute on Aging. Genotyping was funded as part of the NHLBI Candidate-gene Association Resource (N01-HC-65226) and the NHGRI Gene Environment Association Studies (GENEVA) (U01-HG004729, U01-HG04424, and U01-HG004446). This manuscript has been reviewed and approved by CARDIA for scientific content.

The Jackson Heart Study (JHS) is supported by contracts HHSN268201300046C, HHSN268201300047C, HHSN268201300048C, HHSN268201300049C, HHSN268201300050C from the National Heart, Lung, and Blood Institute and the National Institute on Minority Health and Health Disparities. The authors thank the participants and data collection staff of the Jackson Heart Study.

Multi-Ethnic Study of Atherosclerosis (MESA), and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-001079, UL1-TR-000040, and DK063491. The MESA CARe data used for the analyses described in this manuscript were obtained through Genetics (CMP00068). Funding for CARe genotyping was provided by NHLBI Contract N01-HC-65226.

The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

Supplementary Material

Supplementary methods, tables, and figures can be found at http://csb.wfu.edu/SupplementaryData_online.docx.

References

1. Prasad, R. B. & Groop, L. Genetics of type 2 diabetes-pitfalls and possibilities. *Genes* **6,** 87–123 (2015).
2. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44,** 981–990 (2012).

3. Cordell, H. J. Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.* **10,** 392–404 (2009).
4. Moore, J. H. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered.* **56,** 73–82 (2003).
5. Henkin, L. *et al.* Genetic epidemiology of insulin resistance and visceral adiposity. The IRAS Family Study design and methods. *Ann. Epidemiol.* **13,** 211–217 (2003).
6. Pacini, G. & Bergman, R. N. MINMOD: a computer program to calculate insulin sensitivity and pancreatic responsivity from the frequently sampled intravenous glucose tolerance test. *Comput. Methods Programs Biomed.* **23,** 113–122 (1986).
7. The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am. J. Epidemiol.* **129,** 687–702 (1989).
8. Friedman, G. D. *et al.* CARDIA: study design, recruitment, and some characteristics of the examined subjects. *J. Clin. Epidemiol.* **41,** 1105–1116 (1988).
9. Taylor, H. A. *et al.* Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. *Ethn. Dis.* **15,** S6-4–17 (2005).
10. Bild, D. E. *et al.* Multi-ethnic study of atherosclerosis: objectives and design. *Am. J. Epidemiol.* **156,** 871–881 (2002).
11. McDonough, C. W. *et al.* A genome-wide association study for diabetic nephropathy genes in African Americans. *Kidney Int.* **79,** 563–572 (2011).
12. Palmer, N. D. *et al.* A genome-wide association search for type 2 diabetes genes in African Americans. *PloS One* **7,** e29202 (2012).
13. Hellwege, J. N. *et al.* Genome-wide family-based linkage analysis of exome chip variants and cardiometabolic risk. *Genet. Epidemiol.* **38,** 345–352 (2014).
14. O'Connell, J. R. & Weeks, D. E. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am. J. Hum. Genet.* **63,** 259–266 (1998).
15. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
16. Lettre, G. *et al.* Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project. *PLoS Genet.* **7,** e1001300 (2011).
17. Ng, M. C. Y. *et al.* Transferability and fine mapping of type 2 diabetes loci in African Americans: the Candidate Gene Association Resource Plus Study. *Diabetes* **62,** 965–976 (2013).
18. Hester, J. M. *et al.* Implication of European-derived adiposity loci in African Americans. *Int. J. Obes. 2005* **36,** 465–473 (2012).
19. Palmer, N. D. *et al.* Evaluation of DLG2 as a positional candidate for disposition index in African-Americans from the IRAS Family Study. *Diabetes Res. Clin. Pract.* **87,** 69–76 (2010).
20. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).
21. Patterson, N. *et al.* Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74,** 979–1000 (2004).
22. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).

23. Keene, K. L. *et al.* Exploration of the utility of ancestry informative markers for genetic association studies of African Americans with type 2 diabetes and end stage renal disease. *Hum. Genet.* **124,** 147–154 (2008).
24. Almasy, L. & Blangero, J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **62,** 1198–1211 (1998).
25. Qi, L. *et al.* Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. *Hum. Mol. Genet.* **19,** 2706–2715 (2010).
26. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40,** 638–645 (2008).
27. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46,** 234–244 (2014).
28. Voight, B. F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42,** 579–589 (2010).
29. Sim, X. *et al.* Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia. *PLoS Genet.* **7,** e1001363 (2011).
30. Kim, D. *et al.* CXCL12 secreted from adipose tissue recruits macrophages and induces insulin resistance in mice. *Diabetologia* **57,** 1456–1465 (2014).
31. Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* **42,** 105–116 (2010).
32. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk–associated variants. *Nat. Genet.* **46,** 136–143 (2014).
33. Bell, J. T. *et al.* Genome-wide association scan allowing for epistasis in type 2 diabetes. *Ann. Hum. Genet.* **75,** 10–19 (2011).
34. Herold, C., Steffens, M., Brockschmidt, F. F., Baur, M. P. & Becker, T. INTERSNP: genome-wide interaction analysis guided by a priori information. *Bioinforma. Oxf. Engl.* **25,** 3275–3281 (2009).
35. Becker, T., Herold, C., Meesters, C., Mattheisen, M. & Baur, M. P. Significance levels in genome-wide interaction analysis (GWIA). *Ann. Hum. Genet.* **75,** 29–35 (2011).
36. Babron, M.-C., Etcheto, A. & Dizier, M.-H. A New Correction for Multiple Testing in Gene-Gene Interaction Studies. *Ann. Hum. Genet.* (2015). doi:10.1111/ahg.12113
37. Lucas, G. *et al.* Hypothesis-Based Analysis of Gene-Gene Interactions and Risk of Myocardial Infarction. *PLoS ONE* **7,** (2012).

DEEP MOTIF DASHBOARD: VISUALIZING AND UNDERSTANDING GENOMIC SEQUENCES USING DEEP NEURAL NETWORKS

JACK LANCHANTIN, RITAMBHARA SINGH, BEILUN WANG, YANJUN QI

Department of Computer Science, University of Virginia
Charlottesville, VA 22903, USA
E-mail: {jjl5sw,rs3zz,bw4mw,yq2h}@virginia.edu

Deep neural network (DNN) models have recently obtained state-of-the-art prediction accuracy for the transcription factor binding (TFBS) site classification task. However, it remains unclear how these approaches identify meaningful DNA sequence signals and give insights as to why TFs bind to certain locations. In this paper, we propose a toolkit called the Deep Motif Dashboard (DeMo Dashboard) which provides a suite of visualization strategies to extract motifs, or sequence patterns from deep neural network models for TFBS classification. We demonstrate how to visualize and understand three important DNN models: convolutional, recurrent, and convolutional-recurrent networks. Our first visualization method is finding a test sequence's saliency map which uses first-order derivatives to describe the importance of each nucleotide in making the final prediction. Second, considering recurrent models make predictions in a temporal manner (from one end of a TFBS sequence to the other), we introduce temporal output scores, indicating the prediction score of a model over time for a sequential input. Lastly, a class-specific visualization strategy finds the optimal input sequence for a given TFBS positive class via stochastic gradient optimization. Our experimental results indicate that a convolutional-recurrent architecture performs the best among the three architectures. The visualization techniques indicate that CNN-RNN makes predictions by modeling both motifs as well as dependencies among them.

1. Introduction

In recent years, there has been an explosion of deep learning models which have lead to groundbreaking results in many fields such as computer vision,[1] natural language processing,[2] and computational biology.[3–8] However, although these models have proven to be very accurate, they have widely been viewed as "black boxes" due to their complexity, making them hard to understand. This is particularly unfavorable in the biomedical domain, where understanding a model's predictions is extremely important for doctors and researchers trying to use the model.

Aiming to open up the black box, we present the "Deep Motif Dashboard[a]" (DeMo Dashboard), to understand the inner workings of deep neural network models for a genomic sequence classification task. We do this by introducing a suite of different neural models and visualization strategies to see which ones perform the best and understand how they make their predictions.[b]

Understanding genetic sequences is one of the fundamental tasks of health advancements due to the high correlation of genes with diseases and drugs. An important problem within genetic sequence understanding is related to transcription factors (TFs), which are regulatory proteins that bind to DNA. Each different TF binds to specific transcription factor binding sites (TFBSs) on the genome to regulate cell machinery. Given an input DNA sequence, classifying whether or not there is a binding site for a particular TF is a core task of bioinformatics.[10]

For our task, we follow a two step approach. First, given a particular TF of interest and a dataset containing samples of positive and negative TFBS sequences, we construct three deep learning architectures to classify the sequences. Section 2 introduces the three different DNN structures that we use: a convolutional neural network (**CNN**), a recurrent neural network

[a] Dashboard normally refers to a user interface that gives a current summary, usually in graphic, easy-to-read form, of key information relating to performance.[9]

[b] We implemented our model in Torch, and it is made available at deepmotif.org

(**RNN**), and a convolutional-recurrent neural network (**CNN-RNN**).

Once we have our trained models to predict binding sites, the second step of our approach is to understand why the models perform the way they do. As explained in section 3, we do this by introducing three different visualization strategies for interpreting the models:

(1) Measuring nucleotide importance with **Saliency Maps**.
(2) Measuring critical sequence positions for the classifier using **Temporal Output Scores**.
(3) Generating class-specific motif patterns with **Class Optimization**.

We test and evaluate our models and visualization strategies on a large scale benchmark TFBS dataset. Section 4 provides experimental results for understanding and visualizing the three DNN architectures. We find that the CNN-RNN outperforms the other models. From the visualizations, we observe that the CNN-RNN tends to focus its predictions on the traditional motifs, as well as modeling long range dependencies among motifs.

2. Deep Neural Models for TFBS Classification

TFBS Classification. Chromatin immunoprecipitation (ChIP-seq) technologies and databases such as ENCODE[11] have made binding site locations available for hundreds of different TFs. Despite these advancements, there are two major drawbacks: (1) ChIP-seq experiments are slow and expensive, (2) although ChIP-seq experiments can find the binding site locations, they cannot find patterns that are common across all of the positive binding sites which can give insight as to why TFs bind to those locations. Thus, there is a need for large scale computational methods that can not only make accurate binding site classifications, but also identify and understand patterns that influence the binding site locations.

In order to computationally predict TFBSs on a DNA sequence, researchers initially used consensus sequences and position weight matrices to match against a test sequence.[10] Simple neural network classifiers were then proposed to differentiate positive and negative binding sites, but did not show significant improvements over the weight matrix matching methods.[12] Later, SVM techniques outperformed the generative methods by using k-mer features,[13,14] but string kernel based SVM systems are limited by expensive computational cost proportional to the number of training and testing sequences. Most recently, convolutional neural network models have shown state-of-the-art results on the TFBS task and are scalable to a large number of genomic sequences,[3,7] but it remains unclear which neural architectures work best.

Deep Neural Networks for TFBSs. To find which neural models work the best on the TFBS classification task, we examine several different types of models. Inspired by their success across different fields, we explore variations of two popular deep learning architectures: convolutional neural networks (CNNs), and recurrent neural networks (RNNs). CNNs have dominated the field of computer vision in recent years, obtaining state-of-the-art results in many tasks due to their ability to automatically extract translation-invariant features. On the other hand, RNNs have emerged as one of the most powerful models for sequential data tasks such as natural language processing due to their ability to learn long range dependencies. Specifically, on the TFBS prediction task, we explore three distinct architectures: (1) CNN, (2) RNN, and (3) a combination of the two, CNN-RNN. Figure 1 shows an overview of the models.

End-to-end Deep Framework. While the body of the three architectures we use differ, each implemented model follows a similar end-to-end framework which we use to easily compare and contrast results. We use the raw nucleotide characters (A,C,G,T) as inputs, where each character is converted into a one-hot encoding (a binary vector with the matching character entry being a 1 and the rest as 0s). This encoding matrix is used as the input to a convolutional,

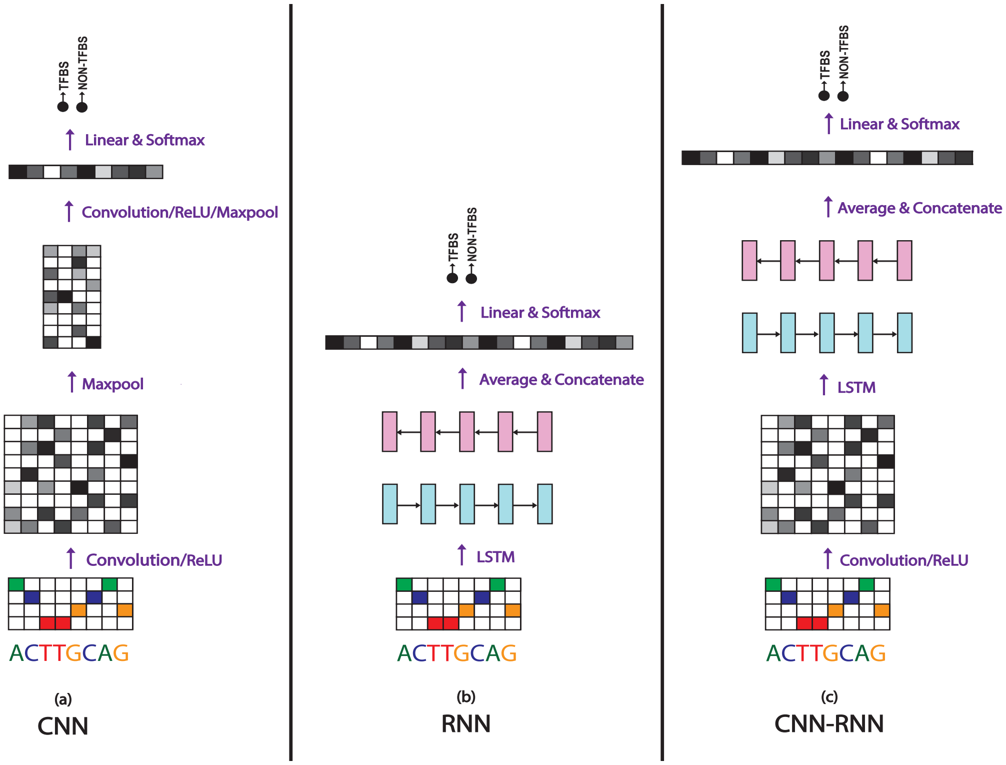


Fig. 1. **Model Architectures.** Each model has the same input (one-hot encoded matrix of the raw nucleotide inputs), and the same output (softmax classifier to make a binary prediction). The architectures differ by the middle "module", which are **(a)** Convolutional, **(b)** Recurrent, and **(c)** Convolutional-Recurrent.

recurrent, or convolutional-recurrent module that each outputs a vector of fixed dimension. The output vector of each model is linearly fed to a softmax function as the last layer which learns the mapping from the hidden space to the output class label space $C \in [+1, -1]$. The final output is a probability indicating whether an input is a positive or a negative binding site (binary classification task). The parameters of the network are trained end-to-end by minimizing the negative log-likelihood over the training set. The minimization of the loss function is obtained via the stochastic gradient algorithm Adam,[15] with a mini-batch size of 256 sequences. We use dropout[16] as a regularization method for each model.

2.1. *Convolutional Neural Network (CNN)*

In genomic sequences, it is believed that regulatory mechanisms such as transcription factor binding are influenced by local sequential patterns known as "motifs". Motifs can be viewed as the temporal equivalent of spatial patterns in images such as eyes on a face, which is what CNNs are able to automatically learn and achieve state-of-the art results on computer vision tasks. As a result, a temporal convolutional neural network is a fitting model to automatically extract these motifs. A temporal convolution with filter (or kernel) size k takes an input data matrix $\mathbf{X}$ of size $T \times n_{in}$, with length T and input layer size n_{in}, and outputs a matrix $\mathbf{Z}$ of size $T \times n_{out}$, where n_{out} is the output layer size. Specifically, $convolution(\mathbf{X}) = \mathbf{Z}$, where

$$\mathbf{z}_{t,i} = \sigma(\mathbf{B}_i + \sum_{j=1}^{n_{in}} \sum_{z=1}^{k} \mathbf{W}_{i,j,z} \mathbf{x}_{t+z-1,j}), \tag{1}$$

where $\mathbf{W}$ and $\mathbf{B}$ are the trainable parameters of the convolution filter, and σ is a function enforcing element-wise nonlinearity. We use rectified linear units (ReLU) as the nonlinearity:

$$\text{ReLU}(x) = \max(0, x). \tag{2}$$

After the convolution and nonlinearity, CNNs typically use maxpooling, which is a dimension reduction technique to provide translation invariance and to extract higher level features from a wider range of the input sequence. Temporal maxpooling on a matrix $\mathbf{Z}$ with a pooling size of m results in output matrix $\mathbf{Y}$. Formally, $maxpool(\mathbf{Z}) = \mathbf{Y}$, where

$$\mathbf{y}_{t,i} = \max_{j=1}^{m} \mathbf{z}_{m(t-1)+j,i} \tag{3}$$

Our CNN implementation involves a progression of convolution, nonlinearity, and maxpooling. This is represented as one convolutional layer in the network, and we test up to 4 layer deep CNNs. The final layer involves a maxpool across the entire temporal domain so that we have a fixed-size vector which can be fed into a softmax classifier.

Figure 1 (a) shows our CNN model with two convolutional layers. The input one-hot encoded matrix is convolved with several filters (not shown) and fed through a ReLU nonlinearity to produce a matrix of convolution activations. We then perform a maxpool on the activation matrix. The output of the first maxpool is fed through another convolution, ReLU, and maxpooled across the entire length resulting in a vector. This vector is then transposed and fed through a linear and softmax layer for classification.

2.2. *Recurrent Neural Network (RNN)*

Designed to handle sequential data, Recurrent neural networks (RNNs) have become the main neural model for tasks such as natural language understanding. The key advantage of RNNs over CNNs is that they are able to find long range patterns in the data which are highly dependent on the ordering of the sequence for the prediction task.

Given an input matrix $\mathbf{X}$ of size $T \times n_{in}$, an RNN produces matrix $\mathbf{H}$ of size $T \times d$, where d is the RNN embedding size. At each timestep t, an RNN takes an input column vector $\mathbf{x_t} \in \mathbb{R}^{n_{in}}$ and the previous hidden state vector $\mathbf{h_{t-1}} \in \mathbb{R}^d$ and produces the next hidden state $\mathbf{h_t}$ by applying the following recursive operation:

$$\mathbf{h}_t = \sigma(\mathbf{W}x_t + \mathbf{U}\mathbf{h}_{t-1} + \mathbf{b}), \tag{4}$$

where $\mathbf{W}, \mathbf{U}, \mathbf{b}$ are the trainable parameters of the model, and σ is an element-wise nonlinearity. Due to their recursive nature, RNNs can model the full conditional distribution of any sequential data and find dependencies over time, where each position in a sequence is a timestep on an imaginary time coordinate running in a certain direction. To handle the "vanishing gradients" problem of training basic RNNs on long sequences, Hochreiter and Schmidhuber[17] proposed an RNN variant called the Long Short-term Memory (LSTM) network (for simplicity, we refer to LSTMs as RNNs in this paper), which can handle long term dependencies by using gating functions. These gates can control when information is written to, read from, and forgotten. Specifically, LSTM "cells" take inputs $\mathbf{x}_t, \mathbf{h}_{t-1}$, and $\mathbf{c}_{t-1}$, and produce $\mathbf{h}_t$, and $\mathbf{c}_t$:

$$\begin{aligned}
\mathbf{i}_t &= \sigma(\mathbf{W}^i\mathbf{x}_t + \mathbf{U}^i\mathbf{h}_{t-1} + \mathbf{b}^i)\\
\mathbf{f}_t &= \sigma(\mathbf{W}^f\mathbf{x}_t + \mathbf{U}^f\mathbf{h}_{t-1} + \mathbf{b}^f)\\
\mathbf{o}_t &= \sigma(\mathbf{W}^o\mathbf{x}_t + \mathbf{U}^o\mathbf{h}_{t-1} + \mathbf{b}^o)\\
\mathbf{g}_t &= tanh(\mathbf{W}^g\mathbf{x}_t + \mathbf{U}^g\mathbf{h}_{t-1} + \mathbf{b}^g)\\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t\\
\mathbf{h}_t &= \mathbf{o}_t \odot tanh(\mathbf{c}_t)
\end{aligned}$$

where $\sigma(\cdot)$, $tanh(\cdot)$, and $\odot$ are element-wise sigmoid, hyperbolic tangent, and multiplication functions, respectively. $\mathbf{i}_t$, $\mathbf{f}_t$, and $\mathbf{o}_t$ are the input, forget, and output gates, respectively.

RNNs produce an output vector $\mathbf{h}_t$ at each timestep t of the input sequence. In order to use them on a classification task, we take the mean of all vectors $\mathbf{h}_t$, and use the mean vector $\mathbf{h}_{mean} \in \mathbb{R}^d$ as input to the softmax layer.

Since there is no innate direction in genomic sequences, we use a bi-directional LSTM as our RNN model. In the bi-directional LSTM, the input sequence gets fed through two LSTM networks, one in each direction, and then the output vectors of each direction get concatenated together in the temporal direction and fed through a linear classifier.

Figure 1 (b) shows our RNN model. The input one-hot encoded matrix is fed through an LSTM in both the forward and backward direction which each produce a matrix of column vectors representing the LSTM output embedding at each timestep. These vectors are then averaged to create one vector for each direction representing the LSTM output. The forward and backward output vectors are then concatenated and fed to the softmax for classification.

2.3. *Convolutional-Recurrent Network (CNN-RNN)*

Considering convolutional networks are designed to extract motifs, and recurrent networks are designed to extract temporal features, we implement a combination of the two in order to find temporal patterns between the motifs. Given an input matrix $\mathbf{X} \in \mathbb{R}^{T \times n_{in}}$, the output of the CNN is $\mathbf{Z} \in \mathbb{R}^{T \times n_{out}}$. Each column vector of $\mathbf{Z}$ gets fed into the RNN one at a time in the same way that the one-hot encoded vectors get input to the regular RNN model. The resulting output of the RNN $\mathbf{H} \in \mathbb{R}^{T \times d}$, where d is the LSTM embedding size, is then averaged across the temporal domain (in the same way as the regular RNN), and fed to a softmax classifier.

Figure 1 (c) shows our CNN-RNN model. The input one-hot encoded matrix is fed through one layer of convolution to produce a convolution activation matrix. This matrix is then input to the LSTM, as done in the regular RNN model from the original one-hot matrix. The output of the LSTM is averaged, concatenated, and fed to the softmax, similar to the RNN.

3. Visualizing and Understanding Deep Models

The previous section explained the deep models we use for the TFBS classification task, where we can evaluate which models perform the best. While making accurate predictions is important in biomedical tasks, it is equally important to understand why models make their predictions. Accurate, but uninterpretable models are often very slow to emerge in practice due to the inability to understand their predictions, making biomedical domain experts reluctant to use them. Consequently, we aim to obtain a better understanding of why certain models work better than others, and investigate how they make their predictions by introducing several visualization techniques. The proposed DeMo Dashboard allows us visualize and understand DNNs in three different ways: Saliency Maps, Temporal Output Scores, and Class Optimizations.

3.1. *Saliency Maps*

For a certain DNA sequence and a model's classification, a logical question may be: "which which parts of the sequence are most influential for the classification?" To do this, we seek to visualize the influence of each position (i.e. nucleotide) on the prediction. Our approach is similar to the methods used on images by Simonyan et al.[18] and Baehrens et al.[19] Given a sequence X_0 of length $|X_0|$, and class $c \in C$, a DNN model provides a score function $S_c(X_0)$. We rank the nucleotides of X_0 based on their influence on the score $S_c(X_0)$. Since $S_c(X)$ is a highly non-linear function of X with deep neural nets, it is hard to directly see the influence of each nucleotide of X on S_c. Mathematically, around the point X_0, $S_c(X)$ can be approximated by a linear function by computing the first-order Taylor expansion:

$$S_c(X) \approx w^T X + b = \sum_{i=1}^{|X|} w_i x_i + b \tag{5}$$

where w is the derivative of S_c with respect to the sequence variable X at the point X_0:

$$w = \left.\frac{\partial S_c}{\partial X}\right|_{X_0} = \textit{saliency map} \tag{6}$$

This derivative is simply one step of backpropagation in the DNN model, and is therefore easy to compute. We do a pointwise multiplication of the saliency map with the one-hot encoded sequence to get the derivative values for the actual nucleotide characters of the sequence (A,T,C, or G) so we can see the influence of the character at each position on the output score. Finally, we take the element-wise magnitude of the resulting derivative vector to visualize how important each character is regardless of derivative direction. We call the resulting vector a "saliency map[18]" because it tells us which nucleotides need to be changed the least in order to affect the class score the most. As we can see from equation 5, the saliency map is simply a weighted sum of the input nucleotides, where the each weight, w_i, indicates the influence of that nucleotide position on the output score.

3.2. *Temporal Output Scores*

Since DNA is sequential (i.e. can be read in a certain direction), it can be insightful to visualize the output scores at each timestep (position) of a sequence, which we call the temporal output scores. Here we assume an imaginary time direction running from left to right on a given sequence, so each position in the sequence is a timestep in such an imagined time coordinate. In other words, we check the RNN's prediction scores when we vary the input of the RNN. The input series is constructed by using subsequences of an input X running along the imaginary time coordinate, where the subsequences start from just the first nucleotide (position), and ends with the entire sequence X. This way we can see exactly where in the sequence the recurrent model changes its decision from negative to positive, or vice versa. Since our recurrent models are bi-directional, we also use the same technique on the reverse sequence. CNNs process the entire sequence at once, thus we can't view its output as a temporal sequence, so we use this visualization on just the RNN and CNN-RNN.

3.3. *Class Optimization*

The previous two visualization methods listed are representative of a specific testing sample (i.e. sequence-specific). Now we introduce an approach to extract a *class-specific* visualization for a DNN model, where we attempt to find the best sequence which maximizes the probability of a positive TFBS, which we call class optimization. Formally, we optimize the following equation where $S_+(X)$ is the probability (or score) of an input sequence X (matrix in our case) being a positive TFBS computed by the softmax equation of our trained DNN model for a specific TF:

$$\arg\max_X S_+(X) + \lambda \|X\|_2^2 \tag{7}$$

where λ is the regularization parameter. We find a locally optimal X through stochastic gradient descent, where the optimization is with respect to the input sequence. In this optimization, the model weights remain unchanged. This is similar to the methods used in Simonyan et al.[18] to optimize toward a specific image class. This visualization method depicts the notion of a positive TFBS class for a particular TF and is not specific to any test sequence.

3.4. *End-to-end Automatic Motif Extraction from the Dashboard*

Our three proposed visualization techniques allow us to manually inspect how the models make their predictions. In order to automatically find patterns from the techniques, we also propose methods to extract motifs, or consensus subsequences that represent the positive binding sites. We extract motifs from each of our three visualization methods in the following ways: (1) From each positive test sequence (thus, 500 total for each TF dataset) we extract a motif from the saliency map by selecting the contiguous length-9 subsequence that achieves the highest sum of contiguous length-9 saliency map values. (2) For each positive test sequence, we extract a motif from the temporal output scores by selecting the length-9 subsequence that shows the strongest score change from negative to positive output score. (3) For each different TF, we can directly use the class-optimized sequence as a motif.

3.5. *Connecting to Previous Studies*

Neural networks have produced state-of-the-art results on several important benchmark tasks related to genomic sequence classification,[3–5] making them a good candidate to use. However, *why* these models work well has been poorly understood. Recent works have attempted to uncover the properties of these models, in which most of the work has been done on understanding image classifications using convolutional neural networks. Zeiler and Fergus[20] used a “deconvolution” approach to map hidden layer representations back to the input space for a specific example, showing the features of the image which were important for classification. Simonyan et al.[18] explored a similar approach by using a first-order Taylor expansion to linearly approximate the network and find the input features most relevant, and also tried optimizing image classes. Many similar techniques later followed to understand convolutional models.[21,22] Most importantly, researchers have found that CNNs are able to extract layers of translational-invariant feature maps, which may indicate why CNNs have been successfully used in genomic sequence predictions which are believed to be triggered by motifs.

On text-based tasks, there have been fewer visualization studies for DNNs. Karpathy et al.[23] explored the interpretability of RNNs for language modeling and found that there exist interpretable neurons which are able to focus on certain language structure such as quotes. Li et al.[24] visualized how RNNs achieve compositionality in natural language for sentiment analysis by visualizing RNN embedding vectors as well as measuring the influence of input words on classification. Both studies show examples that can be validated by our understanding of natural language linguistics. Contrarily, we are interested in understanding DNA “linguistics” given DNNs (the opposite direction of Karpathy et al.[23] and Li et al.[24]).

The main difference between our work and previous works on images and natural language is that instead of trying to understand the DNNs given human understanding of such human perception tasks, we attempt to uncover critical signals in DNA sequences given our understanding of DNNs.

For TFBS prediction, Alipanahi et al.[3] was the first to implement a visualization method on a DNN model. They visualize their CNN model by extracting motifs based on the input subsequence corresponding to the strongest activation location for each convolutional filter (which we call convolution activation). Since they only have one convolutional layer, it is trivial to map the activations back, but this method does not work as well with deeper models. We attempted this technique on our models and found that our approach using saliency maps outperforms it in finding motif patterns (details in section 4). Quang and Xie[4] use the same visualization method on their convolutional-recurrent model for noncoding variant prediction.

Table 1. Variations of DNN Model Hyperparameters

Model	**Conv. Layers**	**Conv. Size (n_{out})**	**Conv. filter Sizes (k)**	**Conv. Pool Size (m)**	**LSTM Layers**	**LSTM Size (d)**
Small RNN	N/A	N/A	N/A	N/A	1	16
Medium RNN	N/A	N/A	N/A	N/A	1	32
Large RNN	N/A	N/A	N/A	N/A	2	32
Small CNN	2	64	9,5	2	N/A	N/A
Medium CNN	3	64	9,5,3	2	N/A	N/A
Large CNN	4	64	9,5,3,3	2	N/A	N/A
Small CNN-RNN	1	64	5	N/A	2	32
Medium CNN-RNN	1	128	9	N/A	1	32
Large CNN-RNN	2	128	9,5	2	1	32

4. Experiments and Results

4.1. *Experimental Setup*

Dataset. In order to evaluate our DNN models and visualizations, we train and test on the 108 K562 cell ENCODE ChIP-Seq TF datasets used in Alipanahi et al.[3] Each TF dataset has an average of 30,819 training sequences (with an even positive/negative split), and each sequence consists of 101 DNA-base characters (A,C,G,T). Every dataset has 1,000 testing sequences (with an even positive/negative split). Positive sequences are extracted from the hg19 genome centered at the reported ChIP-Seq peak. Negative sequences are generated by dinucleotide-preserving shuffle of the positive sequences. Due to the separate train/test data for each TF, we train a separate model for each individual TF dataset.

Variations of DNN Models. We implement several variations of each DNN architecture by varying hyperparameters. Table 1 shows the different hyperparameters in each architecture. We trained many different hyperparameters for each architecture, but we show the best performing model for each type, surrounded by a larger and smaller version to show that it isn't underfitting or overfitting.

Baselines. We use the "MEME-ChIP[25] sum" results from Alipanahi et al.[3] as one prediction performance baseline. These results are from applying MEME-ChIP to the top 500 positive training sequences, deriving five PWMs, and scoring test sequences using the sum of scores using all five PWMs. We also compare against the CNN model proposed in Alipanahi et al.[3] To evaluate motif extraction, we compare against the "convolution activation" method used in Alipanahi et al.[3] and Quang and Xie,[4] where we map the strongest first layer convolution filter activation back to the input sequence to find the most influential length-9 subsequence.

4.2. *TFBS Prediction Performance of DNN Models*

Table 2 shows the mean area under the ROC curve (AUC) scores for each of the tested models (from Table 1). As expected, the CNN models outperform the standard RNN models. This validates our hypothesis that positive binding sites are mainly triggered by local patterns or "motifs" that CNNs can easily find. Interestingly, the CNN-RNN achieves the best performance among the three deep architectures. To check the statistical significance of such comparisons, we apply a pairwise t-test using the AUC scores for each TF and report the two tailed p-values in Table 3. We apply the t-test on each of the best performing (based on AUC) models for each model type. All deep models are significantly better than the MEME baseline. The CNN is significantly better than the RNN and the CNN-RNN is significantly better than the CNN. In order to understand why the CNN-RNN performs the best, we turn to the dashboard visualizations.

Table 2. Mean AUC scores on the TFBS classification task

Model	Mean AUC	Median AUC	STDEV
MEME-ChIP[25]	0.834	0.868	0.127
DeepBind[3] (CNN)	0.903	0.931	0.091
Small RNN	0.860	0.881	106
Med RNN	0.876	0.905	0.116
Large RNN	0.808	0.860	0.175
Small CNN	0.896	0.918	0.098
Med CNN	0.902	0.922	0.085
Large CNN	0.880	0.890	0.093
Small CNN-RNN	0.917	0.943	0.079
Med CNN-RNN	**0.925**	**0.947**	**0.073**
Large CNN-RNN	0.918	0.944	0.081

Table 3. AUC pairwise t-test

Model Comparison[c]	p-value
RNN vs MEME	5.15E-05
CNN vs MEME	1.87E-19
CNN-RNN vs MEME	4.84E-24
CNN vs RNN	5.08E-04
CNN-RNN vs RNN	7.99E-10
CNN-RNN vs CNN	4.79E-22

4.3. *Understanding DNNs Using the DeMo Dashboard*

To evaluate the dashboard visualization methods, we first manually inspect the dashboard visualizations to look for interpretable signals. Figure 2 shows examples of the DeMo Dashboard for three different TFs and positive TFBS sequences. We apply the visualizations on the best performing models of each of the three DNN architectures. Each dashboard snapshot is for a specific TF and contains (1) JASPAR[26] motifs for that TF, which are the "gold standard" motifs generated by biomedical researchers, (2) the positive TFBS class-optimized sequence for each architecture (for the given TF of interest), (3) the positive TFBS test sequence of interest, where the JASPAR motifs in the test sequences are highlighted using a pink box, (4) the saliency map from each DNN model on the test sequence, and (5) forward and backward temporal output scores from the recurrent architectures on the test sequence. In the saliency maps, the more red a position is, the more influential it is for the prediction. In the temporal outputs, blue indicates a negative TFBS prediction while red indicates positive. The saliency map and temporal output visualizations are on the same positive test sequence (as shown twice). The numbers next to the model names in the saliency map section indicate the score outputs of that DNN model on the specified test sequence.

Saliency Maps (middle section of dashboard). By visual inspection, we can see from the saliency maps that CNNs tend to focus on short contiguous subsequences when predicting positive bindings. In other words, CNNs clearly model "motifs" that are the most influential for prediction. The saliency maps of RNNs tend to be spread out more across the entire sequence, indicating that they focus on all nucleotides together, and infer relationships among them. The CNN-RNNs have strong saliency map values around motifs, but we can also see that there are other nucleotides further away from the motifs that are influential for the model's prediction. For example, the CNN-RNN model is 99% confident in its GATA1 TFBS prediction, but the prediction is also influenced by nucleotides outside the motif. In the MAFK saliency maps, we can see that the CNN-RNN and RNN focus on a very wide range of nucleotides to make their predictions, and the RNN doesn't even focus on the known JASPAR motif to make its high confidence prediction.

Table 4. JASPAR motif matches against DeMo Dashboard and baseline motif finding methods using Tomtom

	Saliency Map (out of 500)	Conv. Activations[3,4] (out of 500)	Temporal Output (out of 500)	Class Optimization (out of 57)
CNN	243.9	173.4	N/A	19
RNN	138.6	N/A	53.5	11
CNN-RNN	168.1	74.2	113.2	13

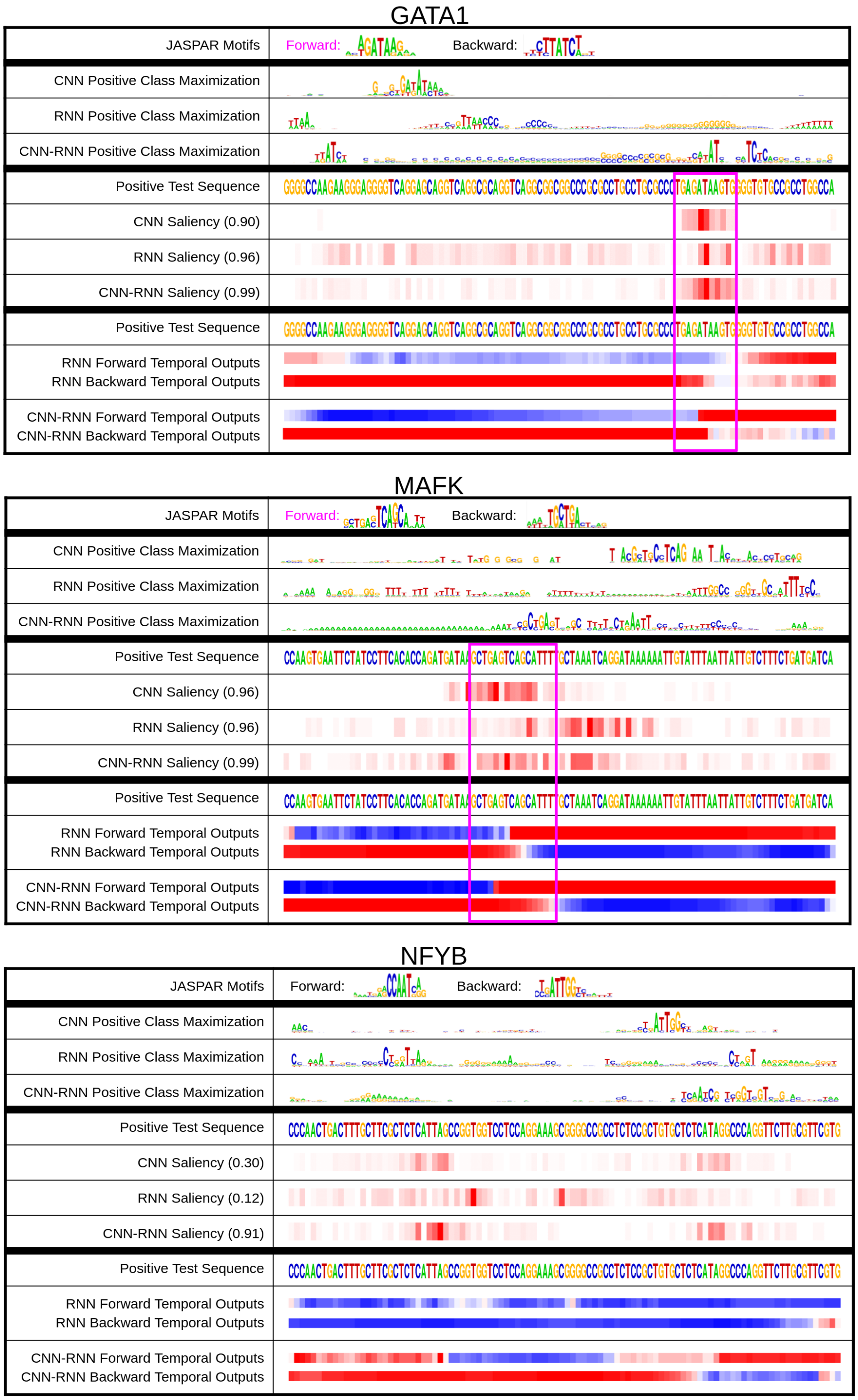


Fig. 2. **DeMo Dashboard**. Dashboard examples for GATA1, MAFK, and NFYB positive TFBS Sequences. The top section of the dashboard contains the Class Optimization (which does not pertain to a specific test sequence, but rather the class in general). The middle section contains the Saliency Maps for a specific positive test sequence, and the bottom section contains the temporal Output Scores for the same positive test sequence used in the saliency map. The very top contains known JASPAR motifs, which are highlighted by pink boxes in the test sequences if they contain motifs.

Temporal Output Scores (bottom section of dashboard). For most of the sequences that we tested, the positions that trigger the model to switch from a negative TFBS prediction to positive are near the JASPAR motifs. We did not observe clear differences between the forward and backward temporal output patterns.

In certain cases, it's interesting to look at the temporal output scores and saliency maps together. An important case study from our examples is the NFYB example, where the CNN and RNN perform poorly, but the CNN-RNN makes the correct prediction. We observe that the CNN-RNN is able to switch its classification from negative to positive, while the RNN never does. To understand why this may have happened, we can see from the saliency maps that the CNN-RNN focuses on two distinct regions, one of which is where it flips its classification from negative to positive. However, the RNN doesn't focus on either of the same areas, and may be the reason why it's never able to classify it as a positive sequence. The fact that the CNN is not able to classify it as a positive sequence, but focuses on the same regions as the CNN-RNN (from the saliency map), may indicate that it is the temporal dependencies between these regions which influence the binding. In addition, the fact that there is no clear JASPAR motif in this sequence may show that the traditional motif approach is not always the best way to model TFBSs.

Class Optimization (top section of dashboard). Class optimization on the CNN model generates concise representations which often resemble the known motifs for that particular TF. For the recurrent models, the TFBS positive optimizations are less clear, though some aspects stand out (like "AT" followed by "TC" in the GATA1 TF for the CNN-RNN). We notice that for certain DNN models, their class optimized sequences optimize the reverse complement motif (e.g. NFYB CNN optimization). The class optimizations can be useful for getting a general idea of what triggers a positive TFBS for a certain TF.

Automatic Motif Extraction from Dashboard. In order to evaluate each DNN's capability to automatically extract motifs, we compare the found motifs of each method (introduced in section 3.4) to the corresponding JASPAR motif, for the TF of interest. We do the comparison using the Tomtom[27] tool, which searches a query motif against a given motif database (and their reverse complements), and returns significant matches ranked by p-value indicating motif-motif similarity. Table 4 summarizes the motif matching results comparing visualization-derived motifs against known motifs in the JASPAR database. We are limited to a comparison of 57 out of our 108 TF datasets by the TFs which JASPAR has motifs for. We compare four visualization approaches: Saliency Map, Convolution Activation,[3,4] Temporal Output Scores and Class Optimizations. The first three techniques are sequence specific, therefore we report the average number of motif matches out of 500 positive sequences (then averaged across 57 TF datasets). The last technique is for a particular TFBS positive class.

We can see from Table 4 that across multiple visualization techniques, the CNN finds motifs the best, followed by the CNN-RNN and the RNN. However, since CNNs perform worse than CNN-RNNs by AUC scores, we hypothesize that this demonstrates that it is also important to model sequential interactions among motifs. In the CNN-RNN combination, CNN acts like a "motif finder" and the RNN finds dependencies among motifs. This analysis shows that visualizing the DNN classifications can lead to a better understanding of DNNs for TFBSs.

5. Conclusions and Future Work

Deep neural networks (DNNs) have shown to be the most accurate models for TFBS classification. However, DNN models are hard to interpret, and thus their adaptation in practice is slow. In this work, we propose the Deep Motif (DeMo) Dashboard to explore three different DNN architectures on TFBS prediction, and introduce three visualization methods to shed light on how these models work. Although our visualization methods still require a human practitioner to examine the dashboard, it is a start to understand these models and we hope that this work will invoke further studies on visualizing and understanding DNN based genomic sequences analysis. Furthermore, DNN models have recently shown to provide excellent results for epigenomic analysis.[8] We plan to extend our DeMo Dashboard to related applications.

References

1. A. Krizhevsky, I. Sutskever and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in *Advances in neural information processing systems*, 2012.
2. I. Sutskever, O. Vinyals and Q. V. Le, Sequence to sequence learning with neural networks, in *Advances in neural information processing systems*, 2014.
3. B. Alipanahi, A. Delong, M. T. Weirauch and B. J. Frey, Predicting the sequence specificities of dna-and rna-binding proteins by deep learning *Nature biotechnology* (Nature Publishing Group, 2015).
4. D. Quang and X. Xie, Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences *bioRxiv* (Cold Spring Harbor Labs Journals, 2015).
5. J. Zhou and O. G. Troyanskaya, Predicting effects of noncoding variants with deep learning-based sequence model *Nature methods* **12** (Nature Publishing Group, 2015).
6. D. R. Kelley, J. Snoek and J. L. Rinn, Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks *Genome research* (Cold Spring Harbor Lab, 2016).
7. J. Lanchantin, R. Singh, Z. Lin and Y. Qi, Deep motif: Visualizing genomic sequence classifications *ICLR Workshops* 2016.
8. R. Singh, J. Lanchantin, G. Robins and Y. Qi, *Bioinformatics* **32**, i639 (2016).
9. Dashboard definiton `http://www.dictionary.com/browse/dashboard`, Accessed: 2016-07-20.
10. G. D. Stormo, Dna binding sites: representation and discovery *Bioinformatics* **16** (Oxford Univ Press, 2000).
11. E. P. Consortium *et al.*, An integrated encyclopedia of dna elements in the human genome *Nature* **489** (Nature Publishing Group, 2012).
12. P. B. Horton and M. Kanehisa, An assessment of neural network and statistical approaches for prediction of e. coli promoter sites *Nucleic Acids Research* **20** (Oxford Univ Press, 1992).
13. M. Ghandi, D. Lee, M. Mohammad-Noori and M. A. Beer, Enhanced regulatory sequence prediction using gapped k-mer features2014.
14. M. Setty and C. S. Leslie, Seqgl identifies context-dependent binding signals in genome-wide regulatory element maps2015.
15. D. Kingma and J. Ba, Adam: A method for stochastic optimization *arXiv preprint arXiv:1412.6980* 2014.
16. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting *The Journal of Machine Learning Research* **15**2014.
17. S. Hochreiter and J. Schmidhuber, Long short-term memory *Neural computation* **9** (MIT Press, 1997).
18. K. Simonyan, A. Vedaldi and A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps *arXiv preprint arXiv:1312.6034* 2013.
19. D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen and K.-R. MÃžller, How to explain individual classification decisions *Journal of Machine Learning Research* **11**2010.
20. M. D. Zeiler and R. Fergus, Visualizing and understanding convolutional networks, in *Computer Vision–ECCV 2014*, (Springer, 2014) pp. 818–833.
21. A. Mahendran and A. Vedaldi, Visualizing deep convolutional neural networks using natural pre-images *International Journal of Computer Vision* (Springer.
22. S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller and W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation *PloS one* **10**2015.
23. A. Karpathy, J. Johnson and F.-F. Li, Visualizing and understanding recurrent networks *arXiv preprint arXiv:1506.02078* 2015.
24. J. Li, X. Chen, E. Hovy and D. Jurafsky, Visualizing and understanding neural models in nlp *arXiv preprint arXiv:1506.01066* 2015.
25. P. Machanick and T. L. Bailey, Meme-chip: motif analysis of large dna datasets *Bioinformatics* **27** (Oxford Univ Press, 2011).
26. A. Mathelier, O. Fornes, D. J. Arenillas, C.-y. Chen, G. Denay, J. Lee, W. Shi, C. Shyr, G. Tan, R. Worsley-Hunt *et al.*, Jaspar 2016: a major expansion and update of the open-access database of transcription factor binding profiles *Nucleic acids research* (Oxford Univ Press, 2015).
27. S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey and W. S. Noble, Quantifying similarity between motifs *Genome biology* **8** (BioMed Central Ltd, 2007).

META-ANALYSIS OF CONTINUOUS PHENOTYPES IDENTIFIES A GENE SIGNATURE THAT CORRELATES WITH COPD DISEASE STATUS

MADELEINE SCOTT*

Stanford University School of Medicine, Stanford University
Stanford, CA 94305, USA

FRANCESCO VALLANIA*

Stanford Institute for Immunity, Transplantation, and Infection, Stanford University
Stanford, CA 94305, USA

PURVESH KHATRI

Stanford Institute for Immunity, Transplantation, and Infection, Stanford University
Dvision of Biomedical Informatics Research, Department of Medicine, Stanford University
Stanford, CA 94305, USA
Email: pkhatri@stanford.edu

**authors contributed equally to this work*

The utility of multi-cohort two-class meta-analysis to identify robust differentially expressed gene signatures has been well established. However, many biomedical applications, such as gene signatures of disease progression, require one-class analysis. Here we describe an R package, MetaCorrelator, that can identify a reproducible transcriptional signature that is correlated with a continuous disease phenotype across multiple datasets. We successfully applied this framework to extract a pattern of gene expression that can predict lung function in patients with chronic obstructive pulmonary disease (COPD) in both peripheral blood mononuclear cells (PBMCs) and tissue. Our results point to a disregulation in the oxidation state of the lungs of patients with COPD, as well as underscore the classically recognized inflammatory state that underlies this disease.

1. Introduction

Chronic obstructive pulmonary disease (COPD) is a progressive, debilitating lung disease that affects one in 20 people across the globe.[1] It is characterized by declining lung function, as measured by Forced Expiratory Volume (FEV) or Global Initiative for Chronic Obstructive Lung Disease (GOLD) stage.[2,3] FEV is the amount of air that a COPD patient can expel in one second, and decreases as the disease progresses. GOLD scoring is the result of a global effort to reach an agreement on spirometric thresholds for COPD diagnosis and is considered the gold standard of COPD severity.[2,3] An increasing GOLD stage reflects declining lung function, where GOLD stage of 0 represents at-risk patients, while a stage of 4 identifies patients with predicted FEV <30%.[3] The rate of COPD progression varies widely from patient to patient, and there are no current treatment options that effectively halt the disease.[4] There is an urgent, critical unmet need to identify pathways that are robustly and reproducibly associated with COPD severity in order to identify novel targets for therapy.

We have previously described a multi-cohort analysis framework for integrated analysis of heterogeneous datasets, and repeatedly demonstrated its successful application across diverse set of diseases including organ transplant, cancer, and infectious diseases for identifying diagnostic, prognostic, and therapeutic signatures.[5–10] At its core, our multi-cohort analysis framework uses random effects inverse variance meta-analysis to identify differentially expressed genes between two groups of samples (e.g., cases vs controls). However, despite its demonstrated utility, its application is limited to two-class comparisons. One of the drawbacks of this framework is that it does not take into account the stage of disease of the patients.[11,12] Further, many biomedical applications, such as those looking to identify signatures of disease progression, require one-class analysis. Such analyses are indispensible for identifying higher risk patients for more personalized care, and to discover pathways associated with disease progression,[12] which in turn could improve our understanding of the disease.

We have implemented an R package, MetaCorrelator, that addresses this challenge and extends the utility of our multi-cohort analysis framework to analyze continuous phenotypes across multiple datasets (Figure 1). MetaCorrelator follows principles of our framework to identify robust signatures for continuous phenotypes. It provides flexibility to use with different continuous phenotypes and widely heterogenous data.

2. Methods

2.1. *Integration of correlation coefficients across independent datasets*

MetaCorrelator starts by computing a correlation coefficient between a designated continuous phenotype and every gene measured in a given discovery dataset. The correlation coefficients can be computed as Pearson's r, Spearman's ρ, or Kendall's τ. Because Spearman's ρ is defined as the Pearson's r calculated on the ranks,[13] it can used directly as r for the rest of the analysis. Kendall's τ need to be converted into r[14] according to

$$r = \sin(\pi * 0.5 * \tau) \tag{1}$$

Then, each correlation coefficient r is converted into a Fisher's Z effect size, defined as:

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r} \tag{2}$$

with variance, V_z, and standard error, SE_z, defined as

$$V_z = \frac{1}{n-3} \tag{3}$$

and

$$SE_z = \sqrt{V_z} \tag{4}$$

where n is the total number of samples used for correlation. Next, we combine Fisher's z for every gene across all discovery datasets into a summary effect size using a random-effects inverse variance model,[15] which assumes that the true effect sizes across each study are not identical but rather sampled from a distribution of true effects. The summary effect size is calculated as

$$Z_s = \frac{\sum_i^n W_i Z_i}{\sum_i^n W_i} \tag{5}$$

and the corresponding summary standard error was computed as

$$SE_s = \sqrt{\frac{1}{\sum_i^n W_i}} \tag{6}$$

where z_i is the Fisher's Z for a given dataset i and W_i is a weight defined as

$$W_i = \frac{1}{V_i + T^2} \tag{7}$$

where V_i is the variance of the Fisher's Z effect size for a given gene within dataset i and T^2 indicates the in-between-dataset variation. Finally, every gene is assigned a p-value calculated using a two-tailed test defined as

$$p = 2[1 - 2(\phi(|\frac{Z_s}{SE_s}|)] \tag{8}$$

The p-value is then corrected for multiple hypothesis testing using Benjamini-Hochberg.

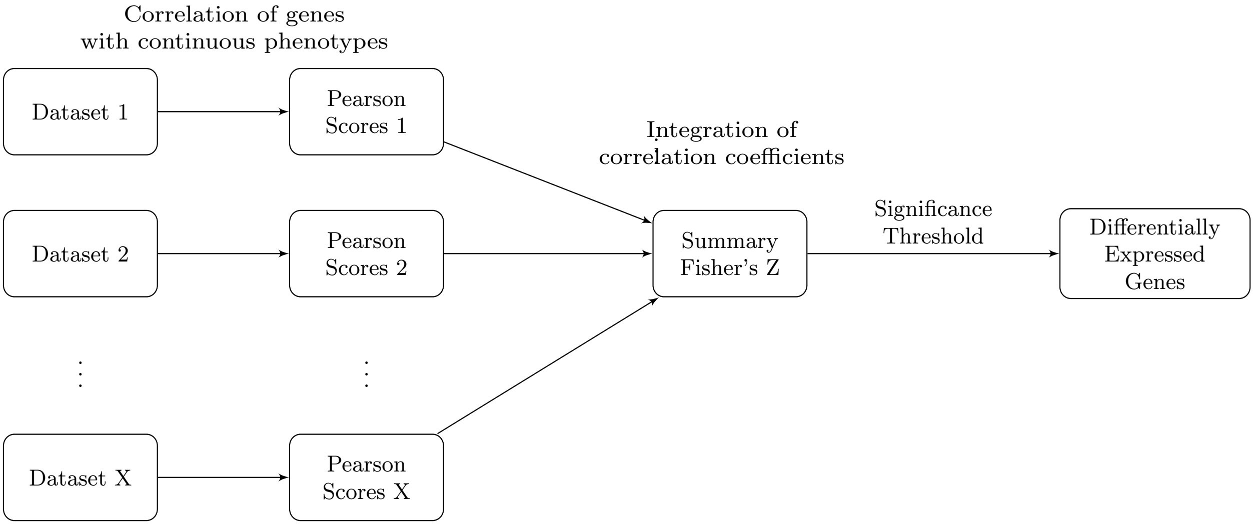

Correlation of genes with continuous phenotypes
Dataset 1
Pearson Scores 1
Dataset 2
Pearson Scores 2
Dataset X
Pearson Scores X
Integration of correlation coefficients
Summary Fisher's Z
Significance Threshold
Differentially Expressed Genes

Fig. 1. Schematic Overview of MetaCorrelator. Each dataset is correlated with a continuous phenotype to compute correlation coefficients (example: Pearson's correlation coefficients). These coefficents are then combined into a summary Fisher's Z effect size. A significance threshold is determined to produce the final list of differentially expressed genes.

2.2. *Datasets*

We identified publicly available gene expression datasets from the NCBI GEO that provided lung function in COPD patients using GOLD stage or FEV. In total, we identified six datasets with 642 samples. All six had expression data that was pre-normalized. We used three datasets for discovery and three as validation (Tables 1 and 2). The datasets were highly heterogeneous; five were from lung biopsies and two from PBMCs, spanning collection over seven years and three countries. All probes were matched to Gene IDs based on the platform information available on GEO. Three of the datasets did not have a control group as all of the samples originated from patients with COPD.

2.3. *Selection and validation of COPD signature*

We used FDR $< 5\%$ to identify genes significantly correlated with COPD severity as defined by GOLD stage or FEV. We performed Gene Ontology enrichment analysis using iPathwayGuide (http://www.advaitabio.com). All other statistical analyses were performed using the statistical programming language R.

2.4. *Availability*

Source code is available at http://khatrilab.stanford.edu/metacorrelator. The Khatri lab will provide full results upon request.

Table 1. Datasets Used in Discovery of Lung Function Signature

GEO ID	Tissue	Phenotype	Cases
GSE47460	Lung Biopsy	GOLD Stage and FEV	75
GSE69818	Lung Biopsy	GOLD Stage	70
GSE76705	PBMCs	FEV	229
3 Datasets	2 Tissues		324

Table 2. Datasets Used to Validate Lung Function Signature

GEO ID	Tissue	Phenotype	Cases
GSE42057	PBMCs	FEV	136
GSE38974	Lung Biopsy	GOLD Stage	32
GSE11906	Lung Biopsy	GOLD Stage	150
3 Datasets	2 Tissues		318

3. Results

3.1. *Functional Analysis of Differentially Expressed Genes Identified by MetaCorrelator*

We identified six independent datasets of 692 lung biopsies or PBMC samples from COPD patients that also provided either GOLD stage or FEV for each patient. The samples included in these datasets came from patients across all stages of COPD and covered all the lobes in the lung. We selected three datasets composed of 374 PBMC samples or lung biopsies as discovery datasets (Table 1), and the rest as validation datasets (Table 2). We choose three discovery datasets such that they increased heterogeneity in the discovery. Two datasets were from lung tissue, and had annotation describing the GOLD stage of the samples. Of the two lung tissue datasets, GSE698181 had COPD patients with and without emphysema. Although GSE47460 had both GOLD stage and FEV annotation, only GOLD stage was used for discovery of the gene signature.

MetaCorrelator identified 108 genes (FDR $< 25\%$) that are consistently correlated with COPD severity as measured by GOLD stage or FEV in the three discovery datasets. We performed Gene Ontology enrichment analysis (Figure 2) to explore the functions of the identified genes. Our enrichment analysis highlighted the role of oxidative stress in COPD progression. We identified the disulfide oxidoreductase activity pathway as a highly significant in COPD progression. This is consistent with previous literature that has identified oxidative stress as a sign of COPD progression.[16]

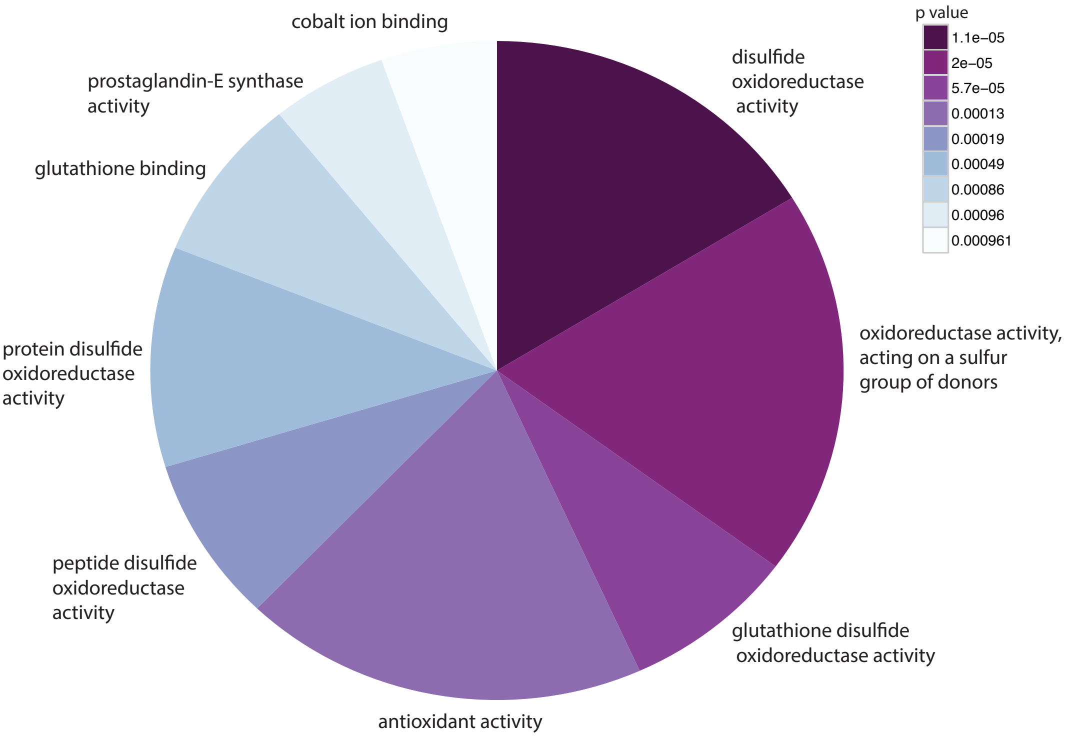


Fig. 2. Functional categories enriched in genes identified by MetaCorrelator

3.2. *Identification and Analysis of a 25-Gene Signature Correlates with COPD Progression*

It is difficult to translate 108-gene signature into a clinical practice. Therefore, to reduce the number of genes, we increased the stringency of our selection criteria by reducing the FDR to 5e-5. We identified 25 genes (7 over-expressed, 18 under-expressed) that are significantly correlated with the COPD severity in the discovery datasets. Enrichment analysis using Gene Ontology of the 25 genes identified matrix remodeling and inflammation as pathways associated with the progression of COPD. Specifically, a subset of the 25 genes, including UDP-Glucuronate Decarboxylase 1 (*UXS1*) and Tetraspanin 13 (*TSPAN13*) are underexpressed genes known to be involved in ECM production and cell adhesion. Importantly, a double tetraspanin knockout mouse (Tetraspanin 28 and 29) has been shown to develop a COPD-like phenotype, underscoring the importance of this protein family to healthy lung function.[18] Inflammatory mediators such as *TLR2* and *FKBP5* are over-expressed in the gene signature, reflecting the inflammatory state of COPD. Previous literature has shown that *TLR2* is over-expressed on Lung CD8+ and CD4+ T-cells as well as CD8+NK T-cells, demonstrating that our results reflect validated biology.[19] Two differentially expressed genes, *TNIK* and *PTPRK* are involved in both ECM and inflammation. Taken together, our results align with previously published literature, which describes COPD as a disregulation of the immune system and subsequent breakdown of the ECM.[20–22]

Next, we defined Lung Function Score (LFS) for a sample as the geometric mean of the expression value of the 25 genes as previously described.[5–7] We observed strong significant correlation between our signature score and FEV in GSE76705 (r = -0.50; p-value = 5.81e-14) (Figure 3). In datasets where GOLD staging was available, we observed a significant score increase in concordance with increasing GOLD stage (JT test; GSE47460: p-value = 9.4e-10; GSE69818: p-value = 5e-4). Interestingly, although only the GOLD stages in GSE47460 were used in the discovery, the LFS strongly correlated with FEV score in GSE47460 (r = -0.57; p-value = 6.23e-4).

3.3. *Validation of the Gene Signature in Three Independent Cohorts*

We validated the 25-gene LFS in three independent cohorts of 318 lung biopsy and PBMC samples from COPD patients (Table 2, Figure 4). Across all three validation cohorts, the LFS was significantly correlated with lung function in the COPD patients (summary effect size = 0.46, p = 3.98e-3). In individual datasets, we observed a significant negative correlation between FEV and LFS in GSE42057 (r = -0.41; p-value = 5.03e-7) and a positive significant correlation with GOLD stages in our remaining two independent cohorts (GSE38974; p-value = 5.539e-6; GSE11906; p-value = 0.02514).

3.4. *Differential Expression in Current vs Never Smokers*

To explore the broader implications of our results, we examined whether any of our 25 identified genes were also significantly expressed in smokers compared to healthy controls. We downloaded seven publicly available datasets from NCBI GEO for a total of 200 samples from smokers and 158 from never smokers (GSE11952, GSE17913, GSE19667, GSE3320, GSE5056, GSE5057, GSE5059). Using the 25-gene LFS derived from MetaCorrelator, two genes, *TSPAN13* and *NR3C2*, were found to be differentially expressed in smokers compared to non-smokers with p value < 0.01. The tetraspanin family has been shown to be critical to normal lung function, and *NR3C2* has been implicated in lung morphogenesis.[23] These results demonstrate the flexibilty of MetaCorrelator to highlight patterns of biological relevance in conjunction with two-class analysis.

4. Discussion

Availability of large amounts of heterogeneous molecular data has necessitated the development of new frameworks to identify patterns and extract new information from these data. We have repeatedly shown the effectiveness of our multi-cohort analysis framework for diagnostic and therapeutic applications across a broad spectrum of human diseases.[5–10] However, this framework is limited to analysis of case-control experiments, and is not suitable for analysis of one-class quantitative phenotypes. Here, we extend our previously established framework to include analysis of gene expression with quantitative quantitative.

Correlation analysis has been a powerful tool for decades, but at this time there does not exist a single framework that can take a collection of datasets and different quantitative

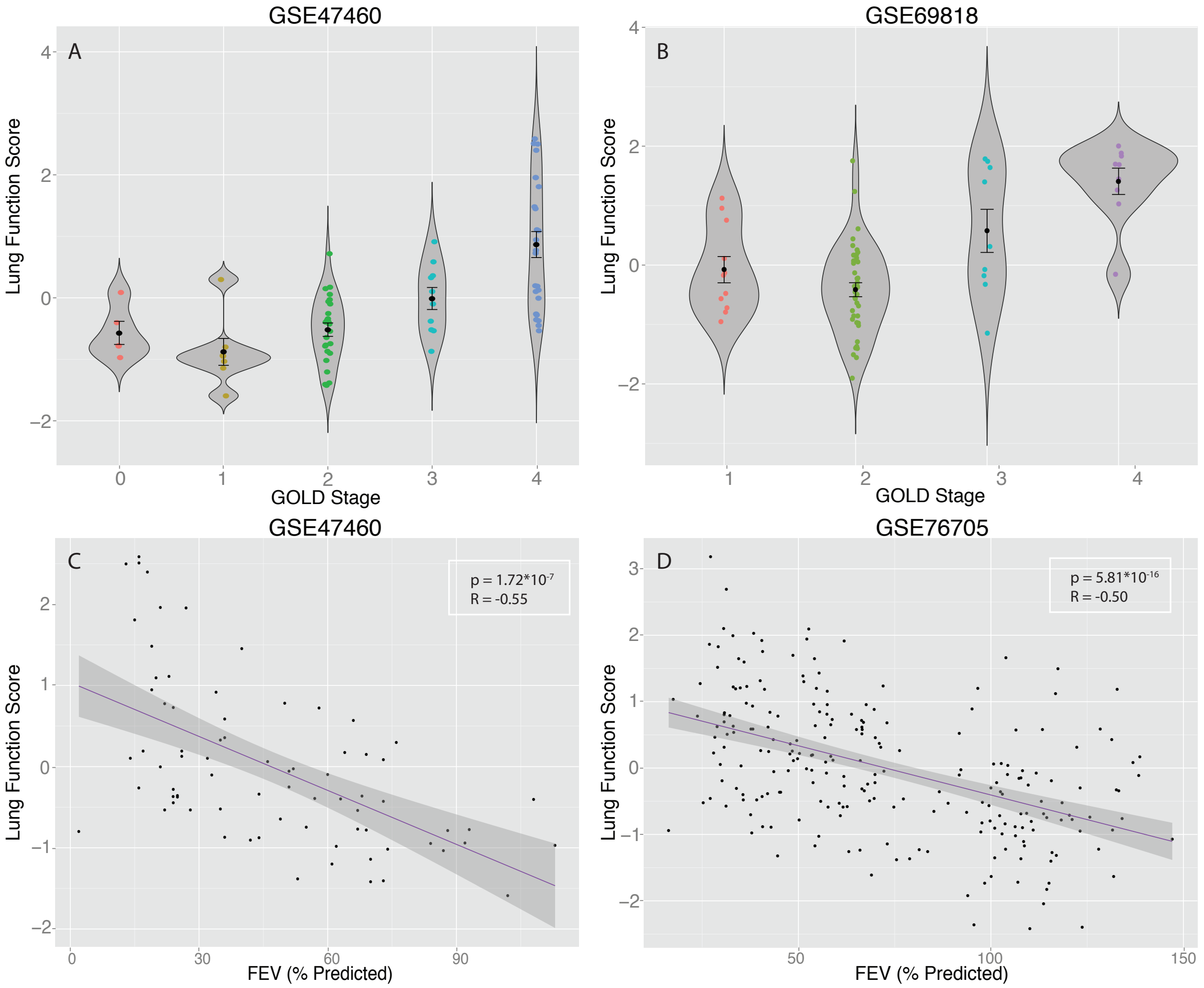


Fig. 3. Lung Function Scores in training cohorts: (A) Violin plots of Lung Function Scores (LFS) in patients distinguished by progressively increasing GOLD stage from GSE47460. (B) Same as (A) but for dataset GSE69818. (C) Correlation plot between LFS and FEV scores in individual patients from GSE47460. (D) Same as (C) but for dataset GSE76705.

phenotypes as input and produce a correlated gene expression signature. There are currently available packages in R, such as metacor, that can compute Fisher's Z values from correlation coefficients; however, MetaCorrelator is uniquely positioned to take multiple datasets as input and correlate gene expression with heterogeneous phenotypes. This is especially relevant in the realm of human disease; methods that are able to integrate different but related organ function phenotypes, such as FEV and GOLD stage, would allow for more powerful analysis that could identify new markers for disease progression.

Our method enables the identification of a gene signature across tissues, thus highlighting the globally relevant differentially expressed genes. By integrating PBMC and lung tissue data, we were able to distill out a gene signature that represents the global differential gene expression of COPD progression. These results emphasize the advantage of integrating multiple tissues. The genes in our signature suggest the importance of inflammation (*TLR2*, *FKBP5*)

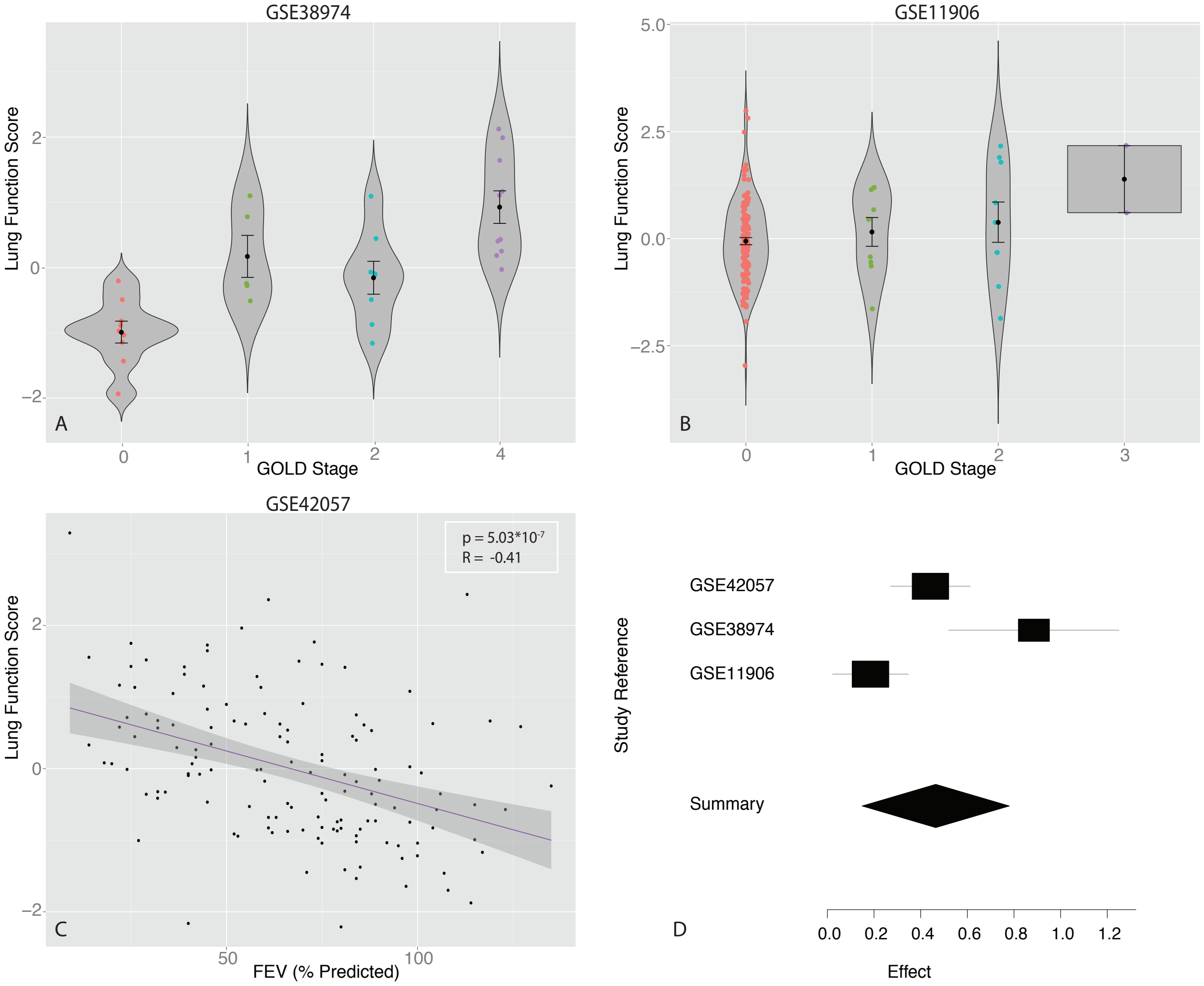


Fig. 4. Lung Function Scores in validation cohorts: (A) Violin plots of Lung Function Scores (LFS) in patients distinguished by progressively increasing GOLD stage from GSE38974. (B) Same as (A) but for dataset GSE11906. (C) Correlation plot between LFS and FEV scores in individual patients from GSE42057. (D) Forest plot representing Fisher's Z values for each of the validation datasets. Squares indicate individual dataset Fisher's Z, with square-size proportional to sample size and horizontal lines indicating individual standard errors (GSE42507 was reverted in sign because of the inverse relationship between GOLD score and FEV). Rhombus indicates summary Fisher's Z with width corresponding to summary standard error.

as well as cell adhesion (*TSPAN13*, *UXS1*), which demonstrates that our framework is able to recapitulate known biology. By integrating the MetaCorrelator framework with established two-class analysis, we can select genes of particular interest. For instance, after identifying differentially expressed genes between smokers and non-smokers, we could further focus on two genes that MetaCorrelator had identified as correlating with COPD progression. MetaCorrelator can be used to correlate any continuous disease phenotype with disease progression. For example, one could identify a gene signature that correlates with prostate specific antigen, a marker of prostate cancer progression. Alternatively, one could correlate a gene signature with ejection fraction of the heart. In summary, MetaCorrelator provides a framework that can correlate whole genome transcriptome across multiple independent datasets with a quantitative phenotype, which in turn can be further explored in case-control studies using the multi-cohort analysis framework.

5. Conclusion

In this study we developed a meta-analysis framework that can integrate multiple gene expression datasets to identify gene signatures that correlate with quantitative phenotypes. Importantly, this method uses the inherent heterogeneity present in multiple cohorts to identify consistently correlated genes and is applicable to datasets that have a single class of sample. Our method can be used in conjunction with other methods that separate samples by class, for example, in order to further differentiate a single group of patients. We applied our method to COPD patients and extracted a 25-gene signature that correlated with lung function in three datasets (two in tissue, one in PBMCs). We then successfully validated our gene signature on three independent datasets. We demonstrated the ability to identify a robust signature with heterogeneous data and phenotypes by correlating the tissue datasets with increasing GOLD stage, and the PBMC dataset with decreasing FEV. Our results suggest an increasing immune response in later stage COPD patients, which has been noted by others, as well as point to a under-appreciated role in sulfur-related oxidative stress. In summary, MetaCorrelator provides a powerful framework to extract a gene signature that is linked to disease progression.

References

1. Halbert, R. J., et al. *European Respiratory Journal* **28**, 523 (2006)
2. Miravitlles, Marc, et al. *Thorax* **64**, 863 (2009).
3. Pauwels, Romain A., et al. *American journal of respiratory and critical care medicine* bf 163, 1256 (2012).
4. Rabe, Klaus F., et al. *American journal of respiratory and critical care medicine* **176**, 532 (2007)
5. Khatri, Purvesh and Roedder, Silke and Kimura, Naoyuki and De Vusser, Katrien and Morgan, Alexander A and Gong, Yongquan and Fischbein, Michael P and Robbins, Robert C and Naesens, Maarten and Butte, Atul J and Sarwal MM. *J. Exp. Med.* **210**, 2205 (2013)
6. M. Andres-Terre, H. M. McGuire, Y. Pouliot, E. Bongen, T. E. Sweeney, C. M. Tato and P. Kha- tri, Immunity **43**, 1199 (December 2015).
7. Timothy E. Sweeney, Aaditya Shidham, Hector R. Wong, and Purvesh Khatri. *Science Translational Medicine* **7**, 287 (2015)
8. Sweeney, Timothy E., Hector R. Wong, and Purvesh Khatri. *Science Translational Medicine* **8** , 346 (2016)
9. R. Chen, P. Khatri, P. K. Mazur, M. Polin, Y. Zheng, D. Vaka, C. D. Hoang, J. Shrager, Y. Xu, S. Vicent, A. J. Butte and E. A. Sweet-Cordero, *Cancer Research* **74**, 2892 (May 2014).
10. T. E. Sweeney, L. Braviak, C. M. Tato and P. Khatri, *The Lancet Respiratory Medicine* **4**, 213 (2016).
11. Walker, Esteban, Adrian V. Hernandez, and Michael W. Kattan. *Cleveland Clinic Journal of Medicine* **75**, 431 (2008)
12. Nordmanna, Alain J., Benjamin Kasendaa, and M. Briel. *Swiss Med Wkly* **142**, w13518 (2012)
13. Myers, Jerome L.; Well, Arnold D. *Research Design and Statistical Analysis* (Routledge, New York, 2003)
14. David A. Walker *JMASM* **2** 525 (2003)
15. M. Borenstein, L.V. Hedges, J.P.T Higgins, and H.R. Rothstein. *Introduction to Meta-Analysis* (Wiley, UK, 2011)
16. Rahman I, Kinnula VL. *Expert Review of Clinical Pharmacology* **5**, 293 (2012).
17. Avrum Spira, Jennifer Beane, Victor Pinto-Plata, Aran Kadar, Gang Liu, Vishal Shah, Bartolome Celli, and Jerome S. Brody. *American Journal of Respiratory Cell and Molecular Biology*, **31**, 601 (2004)
18. Jin, Yingji, et al. *American Thoracic Society* **42**, 633 (2014)
19. Freeman, Christine M., et al. *Respiratory research* **14** (2013)
20. Chung, K. F., and I. M. Adcock. *European Respiratory Journal* **31**, 1334 (2008)
21. Oudijk, EJ D., et al. *Thorax* **60**, 538 (2005)
22. Zandvoort, Andre, et al. *Respiratory research* **9**, 10.1186/1465-9921-9-83 (2008)

23. Duga, Balazs, et al. *Molecular cytogenetics* **7**, 10.1186/1755-8166-7-36 (2014)

PREDICTIVE MODELING OF HOSPITAL READMISSION RATES USING ELECTRONIC MEDICAL RECORD-WIDE MACHINE LEARNING: A CASE-STUDY USING MOUNT SINAI HEART FAILURE COHORT

KHADER SHAMEER[1,2], KIPP W JOHNSON[1,2], ALEXANDRE YAHI [7], RICCARDO MIOTTO[1,2], LI LI[1,2], DORAN RICKS[3], JEBAKUMAR JEBAKARAN[4], PATRICIA KOVATCH[1,4], PARTHO P. SENGUPTA[5], ANNETINE GELIJNS[8], ALAN MOSKOVITZ[8], BRUCE DARROW[5], DAVID L REICH[6], ANDREW KASARSKIS[1], NICHOLAS P. TATONETTI[7], SEAN PINNEY[5] AND JOEL T DUDLEY[1,2,8*]

1. Department of Genetics and Genomics, Icahn Institute of Genomics and Multiscale Biology 2. Institute of Next Generation Healthcare, Mount Sinai Health System 3. Decision Support, Mount Sinai Health System 4. Mount Sinai Data Warehouse, Icahn Institute of Genomics and Multiscale Biology 5. Zena and Michael A. Wiener Cardiovascular Institute, Icahn School of Medicine at Mount Sinai 6. Department of Anesthesiology, Icahn School of Medicine at Mount Sinai 7. Departments of Biomedical Informatics, Systems Biology and Medicine, Columbia University Medical Center, New York 8. Population Health Science and Policy, Mount Sinai Health System, New York, NY

** Corresponding Author, Email: joel.dudley@mssm.edu*

Reduction of preventable hospital readmissions that result from chronic or acute conditions like stroke, heart failure, myocardial infarction and pneumonia remains a significant challenge for improving the outcomes and decreasing the cost of healthcare delivery in the United States. Patient readmission rates are relatively high for conditions like heart failure (HF) despite the implementation of high-quality healthcare delivery operation guidelines created by regulatory authorities. Multiple predictive models are currently available to evaluate potential 30-day readmission rates of patients. Most of these models are hypothesis driven and repetitively assess the predictive abilities of the same set of biomarkers as predictive features. In this manuscript, we discuss our attempt to develop a data-driven, electronic-medical record-wide (EMR-wide) feature selection approach and subsequent machine learning to predict readmission probabilities. We have assessed a large repertoire of variables from electronic medical records of heart failure patients in a single center. The cohort included 1,068 patients with 178 patients were readmitted within a 30-day interval (16.66% readmission rate). A total of 4,205 variables were extracted from EMR including diagnosis codes (n=1,763), medications (n=1,028), laboratory measurements (n=846), surgical procedures (n=564) and vital signs (n=4). We designed a multistep modeling strategy using the Naïve Bayes algorithm. In the first step, we created individual models to classify the cases (readmitted) and controls (non-readmitted). In the second step, features contributing to predictive risk from independent models were combined into a composite model using a correlation-based feature selection (CFS) method. All models were trained and tested using a 5-fold cross-validation method, with 70% of the cohort used for training and the remaining 30% for testing. Compared to existing predictive models for HF readmission rates (AUCs in the range of 0.6-0.7), results from our EMR-wide predictive model (AUC=0.78; Accuracy=83.19%) and phenome-wide feature selection strategies are encouraging and reveal the utility of such data-driven machine learning. Fine tuning of the model, replication using multi-center cohorts and prospective clinical trial to evaluate the clinical utility would help the adoption of the model as a clinical decision system for evaluating readmission status.

1. Introduction

1.1. *Hospital readmission rates – a bottleneck in delivering high value-high volume precision healthcare*

Precision healthcare aims to ensure every patient receive optimal care throughout the onset, maintenance or recovery phases of a disease. Close coordination between different players in the health system is required to integrate and deliver high-quality care. Patients, providers and the care management team play a pivotal role in delivering low-cost, high value and high volume care for patients with diverse healthcare requirements. Improving the quality of healthcare delivery is a challenging task for providers and an important priority for regulatory agencies. As an attempt to reduce healthcare cost, lower healthcare disparities and increase overall quality of care, healthcare regulatory agencies including Centers for Medicaid and Medicare Services (CMS, https://www.cms.gov/) have proposed the Hospital Readmission Reduction Program (HRRP; See: https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/readmissions-reduction-program.html). Depending on the performance of a given provider (or hospital) with respect to the regional, state and federal performance rankings, penalties are levied on healthcare providers. In response, in order to reduce readmissions providers have used commercial or in-house readmission assessment tools to predict 30-day readmission rates, but the overall readmission rates still remain high in various provider sites. In 2015, 2,592 U. S hospitals out of 5,627 registered hospitals in the country received penalties from the CMS (http://khn.org/news/half-of-nations-hospitals-fail-again-to-escape-medicares-readmission-penalties/) for not effectively tackling readmission rates. Despite decades of research, interventions, operational improvements and systems engineering methods, readmission remains a major challenge for patients, providers and payers alike.

1.2. *Readmission rate assessment directive by CMS*

The CMS (https://www.medicare.gov/hospitalcompare/Data/30-day-measures.html) directive on unplanned readmission grades the results of five diseases, two surgical procedures and a quantitative estimate of hospital-wide readmission rates. The conditions that CMS evaluates for readmission rates include three specific cardiovascular diseases (heart attack, heart failure, and stroke), one respiratory disease (chronic obstructive pulmonary disease) and an infectious disease (pneumonia). The hospital-wide readmission rates assess the readmission status of patients admitted to internal medicine, surgery/gynecology, pulmonary, cardiovascular, and neurology services. Further, the 30-day mortality measures determine death rates associated these services. Implementing data-driven methods that consider all available clinical variables in a hypothesis-free approach could identify new features driving clinical outcomes. Such an approach could also provide insights into mechanistic or operational factors that could improve clinical outcomes [1-4]. Heart failure is one of the first core measures by The Joint Commission to assess hospital quality initiatives as part of National Hospital Inpatient Quality Measures. Achieving the lowest readmission rates possible is thus critical to provide high-quality care and improve quality assessments (See: https://www.jointcommission.org/core_measure_sets.aspx).

1.3. *Improving quality of healthcare delivery and outcomes using EMR-wide phenomic data*

Implementation of precision phenotyping algorithms and development of prescriptive prediction models models using phenomic data could aid in the discovery of new knowledge from biomedical and healthcare big data generated in the hospital setting[5,6]. Mining of phenomic big data enables the identification of new or unknown features or combinatorial features driving clinical outcomes. Electronic medical records (EMR) provide access to clinical phenome data and enable better understanding of various clinical phenotypes and the associated outcomes in a systematic manner. Design, development, and deployment of predictive and prescriptive models using EMR-based methods could help to accelerate stratification of patients at risk for improved care. Deploying validated predictive patterns in a clinical setting could improve the quality of healthcare delivery and may have a positive impact on patient outcomes. Phenomics[7] is a relatively new omics term used to define collectively the measurement of phenotypic characteristics of biological entities that include the physical and biochemical traits of organisms including humans. Human phenomics can benefit by leveraging EMRs as a longitudinal data source for the collection of clinical and health traits. While the data currently available within EMR for building a complete picture of a human phenomic state is limited, it is rapidly improving with the integration of genomic data, sensor data and other non-clinical data elements[3,4]. Phenome-wide association studies (PheWAS) studies aim to understand the role of a genetic variant identified from genome-wide association studies (GWAS) in increasing or decreasing the likelihood of observing other diseases in a case-control cohort. PheWAS studies are now revealing the molecular architecture of the pleiotropic nature of genetic variants in mediating multiple diseases[1,8].

1.4. *Predictive modeling of readmission rates in heart failure and need for improvement*

Heart failure is a heterogeneous condition characterized by progressive inability of the heart to supply sufficient blood to the organs of the body. HF is associated with high degree of morbidity and mortality, and 50% of patients with HF die within five years of diagnosis. Heart failure accounts for 43% of Medicare spending even though this patient population only makes up 14% of all Medicare beneficiaries. Heart failure is the top cause of readmission for the Medicare fee-for-service patient population and costs approximately 38 billion dollars annually. Several attempts have reported on the utility, accuracy and actionability of predictive models to model and predict potential readmission associated with heart failure hospitalization. Previously reported models have been built using clinical variables and covariates such as age, sex, race, socioeconomic factors, body mass index, laboratory measures, biomarkers (e.g. B-type natriuretic peptide levels), comorbidities (e.g. neurological disorders, type II diabetes mellitus, etc.), behavioral factors, functional phenotyping of cardiovascular systems (e.g. left ventricular ejection fraction), discharge follow-ups and medications [9-12]. Some models have used billing and procedural codes extracted from EMR or other hospital administration databases. Continuous hemodynamic monitoring devices have also been used to predict readmission rates [13-15]. The predictive power of such HF readmission models remains weak, with Area Under Curve (AUC) values generally in the range of 0.6-0.7. Such models provide only modest utility for predicting which patients may return to the hospital for readmission. There is an immediate need for tools that may be used at the bedside or as part of discharge disposition planning to assess and minimize risk for readmission. Studies led by Hosseinzadeh et.al[16] leverage claims data to predict all-cause readmissions, and Duggal et.al[17] used EMR-derived clinical and administrative data to predict readmission in the setting of a diabetes cohort. To the best of our knowledge, our study is one of the first attempts to use phenome-wide data to identify novel factors driving readmissions related to congestive heart failure and develop EMR-wide prediction models with orthogonal validation to predict the readmission event.

2. Methods

The Mount Sinai Institutional Review Board approved the study. An author (JJ) act as the honest data broker to ensure PHI and HIPAA adherence during the data management, analytics and machine learning. Data scientists and research scientists in the project received a deidentified database from the Mount Sinai Data Warehouse. All analyses were performed using the deidentified data.

2.1. *Mount Sinai Heart cohort and characteristics of heart failure cohort*

The study cohort consists of a database of 1,068 individuals admitted to Mount Sinai Heart service during the year 2014. The principal diagnosis of heart failure using the CMS directive was used to compile HF patients. Each patient readmitted to any service of Mount Sinai within 30-days after the discharge of an HF primary encounter is defined as a "case". The remainder of patients who did not return to the hospital within 30-days were defined as “controls”. Patients admitted to other locations of Mount Sinai Health System or other hospitals within New York city/state or other states in country were not captured. An author (DR) manually phenotyped the cohort and classified the patients as part of a quality control initiative at Mount Sinai Hospital. As an exploratory study with low case rate, no patient exclusion criteria were applied to the dataset.

2.2. *Clinical data analytics and EMR-wide machine learning*

Data was stored in a MySQL database indexed using a unique hexadecimal identifier associated with the data for the visit about HF. Only data about the primary encounter (admission with HF as primary diagnosis) is employed in the analysis. All figures were generated using Wizard for Mac (http://www.wizardmac.com/) and Weka[18-21]. A Naïve Bayes model is used for machine learning. Exploratory data analyses were performed using Elasticsearch and Kibana (https://github.com/elastic/kibana). All models were independently created using 70% of the dataset for training and 30% of the dataset for testing. Bayesian models were created using features unique to each data element and feature selection was performed using correlation based feature subset selection across two classes. Orthogonal validation of machine

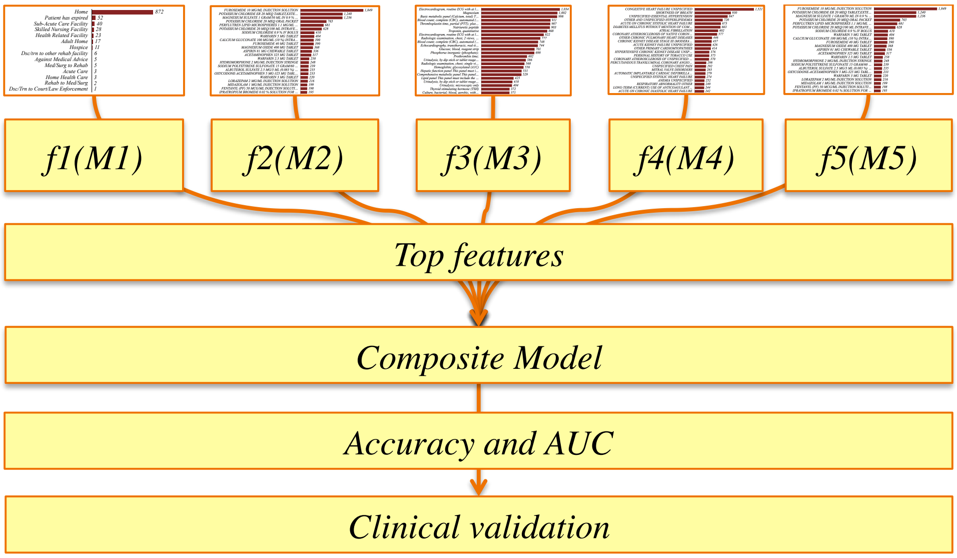


Figure 1: EMR-wide machine learning architecture and predictive modeling strategy to find drivers of readmission rates

learning models was performed with logistic regression. Principal component analyses to understand the variability of features were performed using the Python-based scikit-learn package (http://scikit-learn.org/) and visualized using matplotlib (http://matplotlib.org/). Testing accuracies were estimated using the 5-fold cross validation approach. We define the classification task as a binary classification problem, where RA=”Readmitted” patient and NonRA=”Not readmitted patient”. Weka provides a suite of state-of-the-art machine learning algorithms using a programmatic interface in Java. We used the native Naïve Bayesian classifier in Weka without modification in this exploratory analysis. The algorithm was selected as a rational choice based on prior studies on modeling of readmission prediction[16] Feature ranking and selection[22,23] was performed using a correlation-based feature selection (CFS) method. CFS is a widely used feature selection strategy that aims to find subset of features with significant discriminatory power to perform the classification but which are uncorrelated in feature space. Feature selection is implemented using the “CfsSubsetEval” method in Weka (http://weka.sourceforge.net/doc.dev/weka/attributeSelection/CfsSubsetEval.html). Orthogonal class-specific statistical significance was estimated using Kolmogorov-Smirnov test (distribution estimates), t-test (differences across class-labels), Z-score or Mann-Whitney (median estimates) depending on the data type tested (lab-test, medication, procedure etc.) across the groups (RA and NonRA). An overview of the study design is provided in **Figure 1.**

3. Results

3.1. *Cohort characteristics:*

EMR-wide data mining provides a deep view of various data elements in the cohort (**Figure 2**). A total of 4,205 variables were extracted from EMR. The data from EMR was categorized into five data modalities as diagnosis codes (ICD-9 codes and IMO-codes), procedures (ICD-9, SNOMED-CT and CPT-codes), medications and vital signs. For each patient, the patient encounter specific data is extracted from the EMR. A patient specific filter is used to extract data unique to

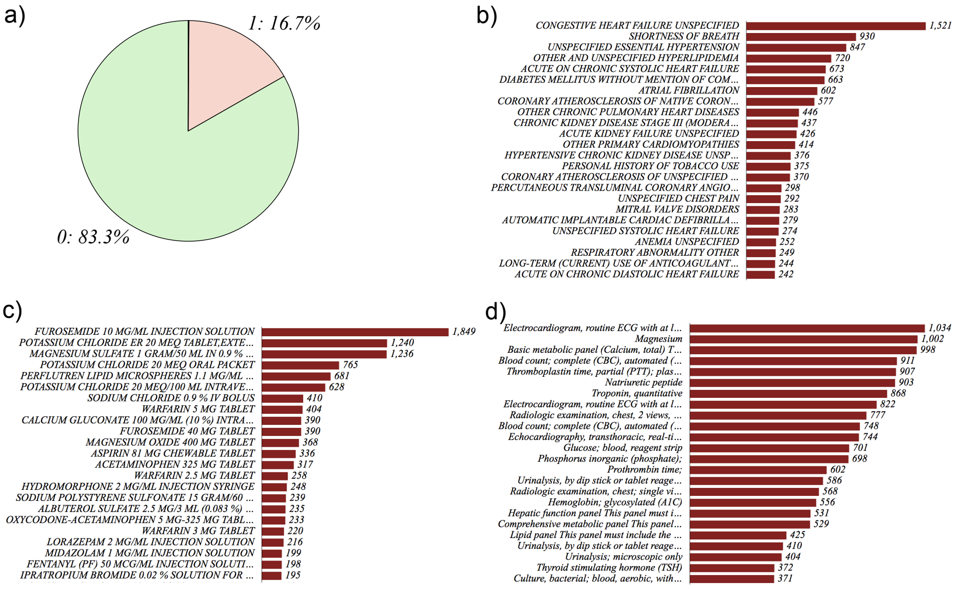


Figure 2: Summary of the study cohort a) case-control ratio: cases are indicated as “1” and controls as “0”. Frequency charts of b) diagnoses c) medications and d) procedures.

the visit; the data from the most recent visit of the patients with multiple admissions is incorporated.

Phenomic data extracted from EMR:

1. Diagnoses codes using ICD-9 (*n*=1,763): ICD-9 codes (http://www.cdc.gov/nchs/icd/icd9.htm) were extracted from Mount Sinai Data Warehouse. The codes were mapped to ICD-9 or IMO codes (https://www.e-imo.com/problemit-terminology-1); all codes were unified to ICD-9 and normalized using UMLS as the bridge (https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html).
2. Medications (n=1,028): Medications prescribed during the hospitalization were compiled using Epic and extracted from Mount Sinai Data warehouse. Medication name, dosage, route of administration was obtained. All medication data was normalized using RxNorm (https://www.nlm.nih.gov/research/umls/rxnorm/).
3. Laboratory measurements (*n*=846): Laboratory measures captured in the EMR were compiled; the raw values of the tests without normalization have been used as a matrix of observations with patients as rows and individual tests as columns.
4. Procedures (*n*=564): Procedures encoded using SNOMED-CT or ICD-9-CM procedures were used.
5. Vital signs (*n*=4): Pulse, respiration rate, systolic blood pressure, heartbeats and temperature were compiled from bedside monitor logs captured in a MySQL database. Vitals were often captured using multiple monitors and approaches. For example temperature was captured at the bedside as axillary temperature, temperature measured via catheter, oral temperature, rectal temperature, or tympanic temperature.

3.2. *EMR-wide feature selection and predictive modeling using five different data modalities*

The machine learning strategy utilized for our study is outlined in Figure 1. To address the trade-offs in dealing with a broad range of features using a small number of samples and missing data, we first generated distinct models using different data elements and relevant features were selected. Features were also compared using orthogonal metrics including logistic regression and PCA to understand the variable space and their inherent relationships. Finally, a composite model for performing predictions is generated using features selected from the individual models. As a real-world machine-learning task, we had a small subset of cases (16.7%) compared to the controls (83.3%). We used a random subset of age and sex matched controls to control the bias introduced by imbalanced datasets. We first generated five different NB predictors using individual data elements. Medications were the most predictive with an accuracy of 81% and AUC of 0.615. Procedure codes encoded as binary variable fared poorly with AUCs of <0.50 (ICD-9 procedures) and 0.553 (CPT codes). We did not generate an independent model for feature selection using the four vital signs after accounting for the small number of features. Laboratory values also showed lower AUC (0.535). Exploration of the data using principal component analyses also revealed that procedures had low variance compared to medications. From a healthcare delivery standpoint, this is insightful, as most of the patients have undergone the same type of procedures in the cardiac units. However the medication profiles of patients may vary due to individualized disease comorbidities, side effect profiles, age, and gender. Details of individual models and features identified using feature selection method (See Table 1). Detailed analyses of medications could provide better insights into features driving readmissions (Johnson & Shameer *et.al*; *manuscript in preparation*)

3.3. *Feature reduction and model refinement*

Due to the low percentage of the cases in the cohort under investigation, a high-dimension feature array is prone to overfitting in machine learning of binary classification tasks. To address this, we have used a feature reduction approach. Features were tested to assess predictive value using a classifier based method and regression models. Feature selection approach and an orthogonal validation approach provide insights into a subset of highly predictive variables associated with readmitted subset of patients. The AUCs of regression models were 0.5685, 0.6471, 0.7596 and 0.795 (ICD-9 and CPT) for vitals, diagnoses codes, medications, and procedures respectively (See Figure 4 and 5). The final composite model is developed using 105 features with an AUC=0.78 and cross-validation testing accuracy of 83.19%.

Data-element	Type	Encoding	Accuracy	AUC	Features
Diagnosis	ICD-9 Diagnosis	Binary	70.3297%	0.605	34/1763
Procedures	ICD-9-Procedure	Binary	77.907%	<0.50	4/273
Procedure	CPT-codes	Binary	72.9858%	0.553	8/564
Medications	Medication name and dosage	Binary	81.9048%	0.615	26/1028
Labs	Non-descriptive lab measurements	Continuous	73.9336%	0.535	29/846
Composite model	Combined features	Hybrid	83.9000%	0.780	105

Table 1: Summary of different Bayesian predictors and features compiled using CFS method

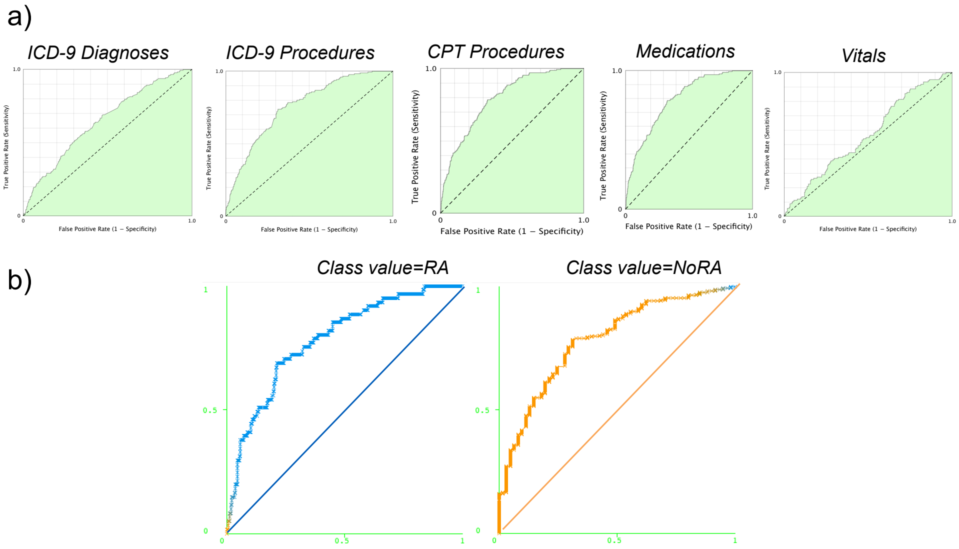


Figure 3: ROC curves a) logistic regression models and b) composite model with 105 features

A brief summary of features significant in feature selection method and the orthogonal validation approach is provided below (also see Figure 5):

a) Procedures: out of 12 procedures, codes for invasive procedures including fine needle aspirations with imaging guidance, intravenous catheterization, routine culture and cell count were significant procedures. As procedures were counted as individual events, the subset of readmitted patients has higher frequency of these procedures compared to patients not readmitted. Repetitive tests for culture and cell count could also indicate potential infection or other complications. *b) Medications:* amongst the 1,028 medications, our analyses indicate 28 medications as features with discriminatory power. Three medications (carvedilol 25 mg tablet, ethacrynic acid IVPB and isosorbide dinitrate 30 mg tablet) were validated using logistic regression approach. However, we noted that only 2.7% of the cohort received carvedilol 25 mg, and all of them were part of the readmission subset. Previous work has potentially indicated that increasing in carvedilol dosage may lead to better a outcome on readmission rate[24]. *c) Diagnosis:* chronic conditions like type 1 diabetes (ICD-9 code 250.01), osteoarthritis; manifestations of cancer (ICD-9 code 233); neurological or psychiatric conditions (mood disorders, hallucinations, sleep disturbances cocaine abuse); cardiovascular structural conditions like rheumatic mitral insufficiency and gastrointestinal conditions such as enteritis were conditions significantly associated with readmission rates. Onco-cardiology assessment of patients may also help in reducing the readmission rates in high-risk patients. Assessment of cardiovascular patients for psychosocial aspects and careful evaluation of individual comorbidities could help to reduce the readmission rates and adherence to the medications [25-28]. *d) Laboratory values:* laboratory values were least predictive in the individual modeling stage. During the orthogonal validation step, creatinine kinase, glucose-fluid, fluid triglycerides and lymphocytes were significant. Optimal glycemic control is a key factor in determining positive outcomes in heart failure patients, especially in those with diabetes mellitus [29]. We noted that features identified using our feature selection method are concordant with earlier findings. For example, we have identified glucose-fluid and type-1 diabetes as predictive factors. We have also identified psychiatric illness, a known factor that influences readmission rates in the setting of complex diseases.

a)

Lab tests	*P*
FLUIDTRIGLYCERIDES	0.007
CK(CPK)	0.01
TOTALPROTEINUR24HR	0.016
U-PROTEIN(CONC.)	0.027
LYMPHOCYTE-PER	0.027
HCGTOTALQUANT.	0.03
GLUCOSEFLUID	0.032
IGGQUANT	0.041
HEMOGLOBINPLASMA	0.048
PROTEIN/CREA.RATIOUR	0.048
ABSNEUTROPHILCOUNT	0.051
5804-PSATOTAL	0.071
1412-U-PROTEIN(RANDOM)	0.102
346-APTT-SL	0.112
UREANITROGEN(POCT)	0.14
1201-OSMOLALITYURINE	0.158
PLATELET	0.18
T4TOTAL	0.246
GLUCOSE(POCT)	0.266
HEMATOCRIT-VEN(ISTAT-MPOCT)	0.336
GLUCOSE(MPOCT)	0.344
NEUTROPHIL-PER	0.387
VANCOMYCINTROUGH	0.63
GAMMAGT	0.743
PROGRAF(FK-506)RANDOM	0.771
NEUTROPHIL	0.809

b)

Procedures	*P*
Cell Count, Body Fluid	0.001
Control of epistaxis by cauterization	0.001
Diabetes mellitus without mention of complication, type I [juvenile type], not stated as uncontrolled	0.001
Rheumatic mitral insufficiency	0.001
Carcinoma in situ of breast and genitourinary system	0.002
Cocaine abuse, continuous	0.002
Culture, Urine, Routine	0.002
Fine needle aspiration with imaging guidance	0.002
Hallucinations	0.002
Insertion of a non-tunneled peripherally inserted central venous catheter, without subcutaneous port	0.002
Osteoarthrosis, unspecified whether generalized or localized, shoulder region	0.002
Other and unspecified episodic mood disorder	0.002
Other Cystoscopy	0.002
Personal history of diseases of blood and blood-forming organs	0.002
Regional enteritis of small intestine with large intestine	0.002
Sleep disturbances	0.002
Pulse	0.321

c)

Medications	*P*
ETHACRYNIC_ACID_IVPB	0.002
ISOSORBIDE_DINITRATE_30_MG_TAB	0.002
CARVEDILOL_25_MG_TABLET	0.015

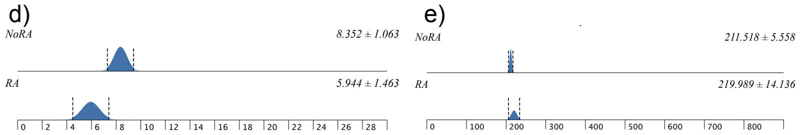


Figure 4: Orthogonal validation of discriminating features a) laboratory tests b) procedures and diagnoses c) medications d) absolute neutrophil count (P=0.051) e) platelet count (P=0.180)

3.4. *Comparison with current heart failure readmission models*

In this work we use EMR-wide feature selection and machine learning to discover novel features and develop new predictors to predict readmission rates. One of the first predictive modeling of hospital readmissions using healthcare data from Quebec, Canada by Hosseinzadeh et.al[16] showed that Naïve Bayes models (0.65) performed better than Random Forest models (0.64). Using a diabetes cohort from a hospital in India, Duggal et.al[17] showed that Naïve Bayes (0.67) showed higher readmission associated savings compared to logistic regression (0.67), Random Forests (0.68), Adaboost (0.67) and Neural Networks (0.62). Futoma et.al[30] showed that Random Forests (0.68) and deep learning using neural networks (0.67) have similar accuracy rate with >1 million patients and > 3 million admission. However, Penalized Logistic Regression had similar accuracy rates as we have shown in our orthogonal validation methods. Compared to existing predictive models for HF readmission rates (AUCs in the range of 0.6-0.7), results from our EMR-wide predictive model (AUC=0.78; Accuracy=83.19%) and phenome-wide feature selection strategies are encouraging and reveal the utility of such data-driven, EMR-wide machine learning.

4. Discussion

Readmission rate is a quality assessment metric routinely used to infer the quality of life index of patient population and the quality of healthcare delivery. Irrespective of the advances in biomedical and healthcare research practices, hospital quality control offices still use traditional readmission risk algorithms and predefined sets of variables to infer the probability patient readmission. However, predictive modeling using big data sourced from different facets of healthcare operations could provide clues to improve the quality of healthcare delivery. Combining predictive analytics with preventive measures would also engage patients, physicians, and payers to participate proactively in improving the health and wellness. Recently we have combined EMR data and genomic data to cluster patients into subtypes with specific genetic variants, disease comorbidities, and medications in a diabetes cohort. Application of deep learning[31,32] in healthcare also shows promise for performing EMR-wide analytics using approaches like Deep Patient[33]. In a recent study, we have created temporal models of disease trajectories that could potentially reveal how the population could cluster into subgroups based on age, gender, self-reported ancestry and comorbidities[34]. Further, we have shown that cognitive machine learning can be utilized for precise phenotyping of high volume echocardiography datasets[35]. We have also applied machine learning to understand various features driving patient satisfaction[36]. Our collective experience in large-scale, automated mining of EMR data suggests that such approaches are useful for both discovery research and the identification of actionable clinical parameters driving diseases or outcomes.

5. Limitations of the current study

In this study, we use all codes without further comprehension; for example, coding systems other than ICD-9 provide an easy way to combine disease. Such an approach could also lead to compiling of similar conditions and hence may not reveal true predictors. For example, we have identified enteritis as a potential diagnosis with readmission. This term would be summarized under gastroenterological conditions. Grouping medication by class or category may also reduce the feature space at the cost of feature resolution. We attempt to capture the best characteristic elements from the real-world data set and hence no data imputation or normalization has been used in our study. The feature selection method may also influence the composition of the models; a systematic assessment of various feature selection algorithms could further enhance the robustness of the model. Healthcare datasets are highly sparse, for example, all patients are not being tested using same laboratory tests except for a few generic tests. Hence, several features may have sparse representations. Even though we had access to EMR-linked genomic data (See BioMe: http://icahn.mssm.edu/research/ipm/programs/biome-biobank), genomic data was not used in this study. Due to a small number of cases; a dramatic increase in feature space would lead to overfitting and high error rates during predictive modeling. We hope to utilize genomic information in a revised version of the model with a larger case dataset. In the current study, we used data from one year of healthcare operations from a single tertiary care healthcare institution. The model should be tested using data from multiple sites and several data-years. Designing of harmonized phenotyping algorithms and data dictionaries leveraging various health information exchanges could help to gather a large number of samples and scale the study using large cohort.

6. Conclusions and Future Directions

A data-driven predictive model is developed to predict readmission rates in heart failure patients. Cases and controls were compiled based on 30-day readmission evidence to the same location. Compared to the existing repertoire of predictive models to assess readmission, our model shows better accuracy using one year of readmission data from a single site. However, the model needs to be updated and calibrated using multiple years of datasets from different sites across the nation. Feature selection provides insights into several novel factors that could help to delineate readmission rates associated with HF. Implementing data-driven methods that EMR-wide variables in a hypothesis-free approach could help us to find new features underlying clinical outcomes. Designing predictive and prescriptive models would help to accelerate stratification of patients at risk for improved care. Such findings and predictive assessments have significant implications for the quality of healthcare delivery and impact on patient outcomes. We envisage that our finding will improve the attempts to develop EMR-wide and scalable phenomics based predictive modeling to find critical events relevant to healthcare delivery and patient outcomes.

7. Acknowledgments

The authors would like to thank the members of the Mount Sinai Health System—Hospital Big Data initiative. This work was supported by a grant from the National Institutes of Health, National Center for Advancing Translational Sciences (NCATS), Clinical and Translational Science Awards (UL1TR001433-01) to KS and JTD.

References

1 Shameer, K. *et al.* A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Hum Genet* **133**, 95-109, doi:10.1007/s00439-013-1355-7 (2014).

2 Glicksberg, B. S. *et al.* An integrative pipeline for multi-modal discovery of disease relationships. *Pac Symp Biocomput*, 407-418 (2015).

3 Badgeley, M. A. *et al.* EHDViz: clinical dashboard development using open-source technologies. *BMJ Open* **6**, e010579, doi:10.1136/bmjopen-2015-010579 (2016).

4 Shameer, K. *et al.* Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. *Brief Bioinform*, doi:10.1093/bib/bbv118 (2016).

5 Hamad, R., Modrek, S., Kubo, J., Goldstein, B. A. & Cullen, M. R. Using "big data" to capture overall health status: properties and predictive value of a claims-based health risk score. *PloS one* **10**, e0126054, doi:10.1371/journal.pone.0126054 (2015).

6 Roski, J., Bo-Linn, G. W. & Andrews, T. A. Creating value in health care through big data: opportunities and policy implications. *Health affairs* **33**, 1115-1122, doi:10.1377/hlthaff.2014.0147 (2014).

7 Houle, D., Govindaraju, D. R. & Omholt, S. Phenomics: the next challenge. *Nat Rev Genet* **11**, 855-866, doi:10.1038/nrg2897 (2010).

8 Karasik, D. How pleiotropic genetics of the musculoskeletal system can inform genomics and phenomics of aging. *Age (Dordr)* **33**, 49-62, doi:10.1007/s11357-010-9159-3 (2011).

9 Thavendiranathan, P. *et al.* Prediction of 30-day heart failure-specific readmission risk by echocardiographic parameters. *Am J Cardiol* **113**, 335-341, doi:10.1016/j.amjcard.2013.09.025 (2014).

10 Padhukasahasram, B., Reddy, C. K., Li, Y. & Lanfear, D. E. Joint impact of clinical and behavioral variables on the risk of unplanned readmission and death after a heart failure hospitalization. *PloS one* **10**, e0129553, doi:10.1371/journal.pone.0129553 (2015).

11 Kansagara, D. *et al.* Risk prediction models for hospital readmission: a systematic review. *JAMA : the journal of the American Medical Association* **306**, 1688-1698, doi:10.1001/jama.2011.1515 (2011).

12 Inouye, S. *et al.* Predicting readmission of heart failure patients using automated follow-up calls. *BMC medical informatics and decision making* **15**, 22, doi:10.1186/s12911-015-0144-8 (2015).

13 Adib-Hajbaghery, M., Maghaminejad, F. & Abbasi, A. The role of continuous care in reducing readmission for patients with heart failure. *J Caring Sci* **2**, 255-267, doi:10.5681/jcs.2013.031 (2013).

14 Bourge, R. C. *et al.* Randomized controlled trial of an implantable continuous hemodynamic monitor in patients with advanced heart failure: the COMPASS-HF study. *J Am Coll Cardiol* **51**, 1073-1079, doi:10.1016/j.jacc.2007.10.061 (2008).

15 Whellan, D. J. *et al.* Development of a method to risk stratify patients with heart failure for 30-day readmission using implantable device diagnostics. *Am J Cardiol* **111**, 79-84, doi:10.1016/j.amjcard.2012.08.050 (2013).

16 Hosseinzadeh, A., Izadi, M., Verma, A., Precup, D. & Buckeridge, D. in *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence* 1532-1538 (AAAI Press, Bellevue, Washington, 2013).

17 Duggal, R., Shukla, S., Chandra, S., Shukla, B. & Khatri, S. K. Predictive risk modelling for early hospital readmission of patients with diabetes in India. *International Journal of Diabetes in Developing Countries*, 1-10, doi:10.1007/s13410-016-0511-8 (2016).

18 Gewehr, J. E., Szugat, M. & Zimmer, R. BioWeka--extending the Weka framework for bioinformatics. *Bioinformatics* **23**, 651-653, doi:10.1093/bioinformatics/btl671 (2007).

19 Hall, M. *et al.* The WEKA Data Mining Software: An Update. *SIGKDD Explor Newsl* **11**, 10-18 (2009).

20 Pyka, M., Balz, A., Jansen, A., Krug, A. & Hullermeier, E. A WEKA interface for fMRI data. *Neuroinformatics* **10**, 409-413, doi:10.1007/s12021-012-9144-3 (2012).

21 Smith, T. C. & Frank, E. Introducing Machine Learning Concepts with WEKA. *Methods Mol Biol* **1418**, 353-378, doi:10.1007/978-1-4939-3578-9_17 (2016).

22 Guyon, I., Andr, #233 & Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157-1182 (2003).

23 Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*. (Springer New York Inc., 2001).

24 Doughty, R. N. & White, H. D. Carvedilol: use in chronic heart failure. *Expert Rev Cardiovasc Ther* **5**, 21-31, doi:10.1586/14779072.5.1.21 (2007).

25 Richardson, L. G. Psychosocial issues in patients with congestive heart failure. *Prog Cardiovasc Nurs* **18**, 19-27 (2003).

26 MacMahon, K. M. & Lip, G. Y. Psychological factors in heart failure: a review of the literature. *Arch Intern Med* **162**, 509-516 (2002).

27 Schweitzer, R. D., Head, K. & Dwyer, J. W. Psychological factors and treatment adherence behavior in patients with chronic heart failure. *J Cardiovasc Nurs* **22**, 76-83 (2007).

28 Ramasamy, R. *et al.* Psychological and social factors that correlate with dyspnea in heart failure. *Psychosomatics* **47**, 430-434, doi:10.1176/appi.psy.47.5.430 (2006).

29 Iribarren, C. *et al.* Glycemic control and heart failure among adult patients with diabetes. *Circulation* **103**, 2668-2673 (2001).

30 Futoma, J., Morris, J. & Lucas, J. A comparison of models for predicting early hospital readmissions. *J Biomed Inform* **56**, 229-238, doi:10.1016/j.jbi.2015.05.016 (2015).

31 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444, doi:10.1038/nature14539 (2015).

32 Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw* **61**, 85-117, doi:10.1016/j.neunet.2014.09.003 (2015).

33 Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep* **6**, 26094, doi:10.1038/srep26094 (2016).

34 Benjamin S. Glicksberg, L. L., Marcus A. Badgeley, Khader Shameer, Roman Kosoy, Noam D. Beckmann, Nam Pho, Jörg Hakenberg, Meng Ma, Kristin L. Ayers, Gabriel E. Hoffman, Shuyu Dan Li, Eric E. Schadt, Chirag J. Patel, Rong Chen, and Joel T. Dudley. Comparative Analyses of Population-scale Phenomic Data in Electronic Medical Records Reveal Race-specific Disease Networks. *Bioinformatics* ISCB Special Issue, doi:10.1093/bioinformatics/btw282 (2016).

35 Sengupta, P. P. *et al.* Cognitive Machine Learning Algorithm for Cardiac Imaging: A Pilot Study for Differentiating Constrictive Pericarditis From Restrictive Cardiomyopathy. *Circ Cardiovasc Imaging* **9**, doi:10.1161/CIRCIMAGING.115.004330 (2016).

36 Li, L., Lee, N. J., Glicksberg, B. S., Radbill, B. D. & Dudley, J. T. Data-Driven Identification of Risk Factors of Patient Satisfaction at a Large Urban Academic Medical Center. *PLoS One* **11**, e0156076, doi:10.1371/journal.pone.0156076 (2016).

LEARNING PARSIMONIOUS ENSEMBLES FOR UNBALANCED COMPUTATIONAL GENOMICS PROBLEMS

ANA STANESCU and GAURAV PANDEY*

Icahn Institute for Genomics and Multiscale Biology and
Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai
New York, NY, USA
**E-mail: gaurav.pandey@mssm.edu*

Prediction problems in biomedical sciences are generally quite difficult, partially due to incomplete knowledge of how the phenomenon of interest is influenced by the variables and measurements used for prediction, as well as a lack of consensus regarding the ideal predictor(s) for specific problems. In these situations, a powerful approach to improving prediction performance is to construct ensembles that combine the outputs of many individual base predictors, which have been successful for many biomedical prediction tasks. Moreover, selecting a *parsimonious* ensemble can be of even greater value for biomedical sciences, where it is not only important to learn an accurate predictor, but also to interpret what novel knowledge it can provide about the target problem. Ensemble selection is a promising approach for this task because of its ability to select a collectively predictive subset, often a relatively small one, of all input base predictors. One of the most well-known algorithms for ensemble selection, CES (Caruana *et al.*'s Ensemble Selection), generally performs well in practice, but faces several challenges due to the difficulty of choosing the right values of its various parameters. Since the choices made for these parameters are usually ad-hoc, good performance of CES is difficult to guarantee for a variety of problems or datasets. To address these challenges with CES and other such algorithms, we propose a novel heterogeneous ensemble selection approach based on the paradigm of reinforcement learning (RL), which offers a more systematic and mathematically sound methodology for exploring the many possible combinations of base predictors that can be selected into an ensemble. We develop three RL-based strategies for constructing ensembles and analyze their results on two unbalanced computational genomics problems, namely the prediction of protein function and splice sites in eukaryotic genomes. We show that the resultant ensembles are indeed substantially more parsimonious as compared to the full set of base predictors, yet still offer almost the same classification power, especially for larger datasets. The RL ensembles also yield a better combination of parsimony and predictive performance as compared to CES.

Keywords: Heterogeneous ensembles; Ensemble selection; Reinforcement learning; Computational genomics.

1. Introduction

Prediction problems in biomedical sciences, such as protein function prediction,[1,2] drug target discovery,[3] and classification of genomic elements[4] are generally quite difficult. This is due in part to incomplete knowledge of how the phenomenon of interest is influenced by the variables and measurements used for prediction, as well as a lack of consensus regarding the ideal predictor(s) for specific problems. From a data perspective, the frequent presence of extreme class imbalance, missing values, heterogeneous data sources of different scales, overlapping feature distributions, and measurement noise further complicate prediction.

In scenarios like these, a powerful approach to improving prediction performance is to construct ensemble predictors that combine the output of many individual base predictors.[5,6] These ensembles have been very successful in producing accurate predictions for many biomedical prediction tasks.[7–13] The success of these methods is attributed to their ability to reinforce accurate predictions as well as correct errors across many diverse base predictors.[14] Diversity among the base predictors is key to ensemble performance: If there is complete consensus (no diversity), the ensemble does not provide any advantage over any of the base predictors. Similarly, an ensemble lacking any consensus (highest diversity) is unlikely to produce confident predictions. Successful ensemble methods strike a balance between diversity and accuracy.[15,16] Popular methods like boosting[17] and random forest[18] generate this diversity by sampling from or assigning weights to training examples. However, they generally utilize a single type of base predictor, such as decision trees, to build the ensemble. Such *homogeneous* ensembles may not be the best choice for problems in biomedical sciences, where the ideal base prediction method is often unclear due to incomplete knowledge and/or data issues.

A more potent approach in this scenario is to build ensembles from the predictions of a wide variety of *heterogeneous* base predictors. Two commonly used heterogeneous ensemble methods are *stacking*,[19,20] and *ensemble selection*.[21,22] Recently, we showed that these methods are more effective than homogeneous ensembles and individual classification methods for complex prediction problems in genomics.[23,24] Other studies have produced similar results.[8,25]

Ensemble selection is an especially promising approach, not only for improving prediction performance, but also because of its ability to select a collectively predictive subset, often a relatively small one, of all input base predictors. This ability to select a *parsimonious* ensemble can be very valuable for biomedical sciences, where it is not only important to learn an accurate predictor, but also to interpret what novel knowledge it can provide about the target problem. For instance, in predicting protein function,[1] it is critical to identify the biological features or principles on the basis of which accurate predictions of protein function are made.[2] It would be easier to reverse engineer a smaller (more parsimonious) ensemble to identify such features or principles than a much larger one, such as all the base predictors taken together. This goal motivated us to develop better algorithms for ensemble selection.

The most well-known algorithm for ensemble selection, which we will refer to as CES (Caruana *et al.*'s Ensemble Selection),[21,22] iteratively grows an ensemble by adding base predictors that produce a gain in prediction performance by (indirectly) enhancing the diversity of the ensemble. Although CES generally performs well in practice, it faces several challenges due to the difficulty of choosing the right values of its various parameters (for details, refer to Section 2.2). For instance, it is unclear how many base predictors should comprise the final ensemble of CES, how many (one or more) should be added in each iteration of the algorithm, and what the right termination condition should be. Since the choices made for these parameters are usually ad-hoc, good performance of CES is difficult to guarantee for a variety of problems or datasets.

To address these challenges with CES and other such algorithms, we propose a novel heterogeneous ensemble selection approach based on the well-established paradigm of reinforcement learning (RL).[26] We demonstrate how RL offers a more systematic and mathematically sound methodology for exploring the many possible combinations of base predictors that can be selected into an ensemble. We test our proposed approaches, and several baselines, on two important computational genomics problem, namely the prediction of protein function and splice sites in eukaryotic genomes. We focus on these unbalanced problems as effective individual (base) predictive models are difficult to learn for them, thus offering a suitable use case for testing ensemble predictors. Our results show that the RL ensembles are indeed substantially more parsimonious with respect to the full set of base predictors, but still offer almost the same predictive power, especially for larger datasets. The RL ensembles also yield a better combination of parsimony and predictive performance as compared to CES. We expect our approaches to be effective for other biomedical problems as well as aid in interpretability, although the latter is a challenging and often subjective task for complex problems. Thus, interpretation of the ensembles we discover is outside the scope of this work.

2. Preliminary Materials and Methods

2.1. *Problem definitions and datasets*

We assess the predictive ability of various ensemble (selection) techniques, such as CES and our RL-based ones, on several protein function (PF) and splice site (SS) prediction datasets.

Protein Function Prediction: Gene expression data are commonly used for predicting protein function, as the simultaneous measurement of gene expression across the entire genome enables effective inference of functional relationships and annotations.[1,2] Thus, for the PFP assessment, we use the gene expression compendium of Hughes *et al.*[27] to predict the functions of roughly 4,000 *S. cerevisiae* genes. Among these genes, the three most abundant functional labels (GO terms) from the list of most biologically interesting and actionable Gene Ontology Biological Process terms compiled by Myers *et al.*[28] are used in our evaluation. These labels are GO:0051252 (regulation of RNA metabolic process), GO:0006366 (transcription from RNA polymerase II promoter) and GO:0016192 (vesicle-mediated transport). We refer to these prediction problems as PF1, PF2, and PF3 respectively (details in Table 1).

Prediction of Splicing Sites: RNA splicing is a naturally occurring phenomenon that contributes to protein diversity. Generally, when creating mature RNA from DNA, introns are removed (or spliced out) from the

Table 1. Details of protein function (PF) and splice site (SS) datasets, including the number of features, number of examples in the minority (positive) and majority (negative) classes, and total number of examples.

	Protein Function Datasets (PF)			Splice Site Datasets (SS)				
Problem	PF1	PF2	PF3					
		(*S. cerevisiae*)		*D. melanogaster*	*C. elegans*	*P. pacificus*	*C. remanei*	*A. thaliana*
#Features	300	300	300	141	141	141	141	141
#Positives	382	344	327	1,598	997	1,596	1,600	1,600
#Negatives	3,597	3,635	3,652	158,150	99,003	156,326	157,542	158,377
#Total	3,979	3,979	3,979	159,748	100,000	157,922	159,142	159,977

gene sequence and exons are retained (or transcribed). Splice sites are conserved nucleotide dimers found at the interfaces between exons and introns. In general, splice sites are *canonical*, as acceptor splice sites are signaled by the occurrence of the consensus dimer “AG” at the $3'$ end of the intron, while donor splice sites are characterized by the consensus dimer “GT”, situated at the $5'$ end of the intron. Such dimers occur frequently throughout most eukaryotic genomes but their presence alone is not sufficient to declare a splice site. Correctly identifying splice sites is an essential step towards genome annotation, and a difficult problem due to the highly unbalanced ratio of bona fide splice sites to decoy dimers.[29] In this work, we focus on identifying acceptor splice sites. In machine learning terms, the problem is formulated as a binary classification of DNA sequences (141-nucleotide-long windows around “AG” dimers, with the dimer situated at position 61) as true acceptor splice sites and decoy dimers. Thus, we assess the ability of ensemble learning to address this important problem on five datasets of acceptor splice sites from five organisms: *D. melanogaster, C. elegans, P. pacificus, C. remanei*, and *A. thaliana*, published by Schweikert *et al.*[4] and Rätsch *et al.*[30]

Note that in some of our experimental evaluations, we investigate the PF and SS datasets results separately due to the substantially different numbers of examples in these datasets.

2.2. *Ensemble selection with CES*

Caruana *et al.*’s Ensemble Selection (CES)[21,22] is arguably the most well-known ensemble selection method. CES begins with an ensemble consisting of the best n individual base predictors (n = 1 in our implementation), and iteratively adds new predictors that maximize the performance of the resultant ensemble on a validation set according to a chosen measure. In each iteration, a pool of m of the candidate base predictors are selected randomly without replacement (m = *#base predictors* in our implementation[22]), and the performance resulting from the addition of each individual candidate to the current ensemble is evaluated. The candidate resulting in the best ensemble performance is selected, the ensemble is updated with it, and the process repeats until a maximum ensemble size (set to the *total #base predictors*[22] in our implementation) is reached. We also varied the values of the pool size (m) and maximum size to test the sensitivity of CES to these parameters.

2.3. *Reinforcement Learning (RL)*

The RL machine learning paradigm[26] allows decision-making software agents to learn the ideal behavior within a specific observable environment such that their performance at the specified task is maximized. In order to learn its behavior, an agent requires feedback for its actions, given by the environment in the form of reinforcement signals. Learning what the best action an agent can take given the current state is achieved through trial-and-error interactions with the environment. One of the most basic applications of RL is teaching a novice robot how to traverse a room from one end to another with various obstacles being placed at random locations in the room; for more complex robotics tasks refer to Kober *et al.*[31]

Generally, RL problems are formulated in terms of Markov Decision Processes (MDPs),[32] a commonly adopted framework for modeling environments and sequential decision making. The environment is made up of a finite set of states S, and the actions come from a discrete set A of actions allowed in a given state. The agent’s job is to investigate the environment by taking actions and observing the rewards. The Markov property affirms that the current state contains enough information to make a decision about the next action. The goal of RL is to repeat the action-observation process that results in the agent learning a good/optimal strategy, called “policy”, for collecting rewards, and completing the task(s) at hand.

Since there is no prior knowledge about any rewards, nor any transition probabilities from any states, (in other words, there is no model available), the agent has to actively "sample" the MDP. Hence the need for *exploration*: the agent tries different actions, often previously untried ones, and assesses their outcome. However, in order to gain sufficient cumulative reward, the agent must also *exploit* its current knowledge about actions already tested that have proved to be beneficial. This exploration/exploitation balance is critical, yet difficult to determine, and represents a fundamental dilemma in RL scenarios. This balance is usually enforced by the classic ϵ-greedy strategy:[33] with probability ϵ, the agent takes a randomly selected action (exploration), and with probability $1 - \epsilon$, the agent chooses the action with the highest estimated payoff (exploitation).

2.4. *Q-Learning*

Following the ϵ-greedy exploration mechanism described above, the agent is able to gather enough information about the environment and create its own model, more specifically, to estimate the quality of each state-action combinations; this mapping is known as the Q-value function. The agent takes an action a_t in a state s_t, ends up in state s_{t+1}, and receives a reward r_t; subsequently, the Q-value associated with action a_t in state s_t is updated. One of the most popular ways for estimating Q-value functions in a model-free framework is the Q-learning algorithm.[34] Under specific conditions/assumptions, Q-learning is able to find an optimal policy (π) regardless of the model the agent adopts, *i.e.*, which action a_t it takes in state state s_t, provided it tries all actions of all states infinitely many times. The policy, which is the agent's learned way of behaving in the environment, is estimated by continuously updating the action-value function $Q(s_t, a_t)$, as described in Equation 1. Here, the learning rate α controls how quickly the learning occurs, and the discount factor γ controls how important future rewards are.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \cdot \left(r_{t+1} + \gamma \cdot \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right) \tag{1}$$

3. Proposed Approach

Although CES (Section 2.2) represents an intuitive and straight-forward approach to ensemble selection, several of its inputs/parameters are very difficult to specify beforehand, which is very likely to impact its performance adversely. For instance, the best values of parameters m and n are difficult to specify a priori for a given dataset/problem. Similarly, the termination criteria, *i.e.*, when the CES ensemble should stop growing so as to avoid overfitting, are also often unclear, and are often set in an ad-hoc manner. Due to these factors, CES is only able to perform an ad-hoc sub-optimal search of the possible ensemble space, and not an optimized exhaustive one. We address these challenges of CES and make ensemble selection more rigorous and exhaustive by leveraging concepts from RL (Section 2.3). In particular, we take advantage of the Q-learning algorithm (Section 2.4), which is proven to converge to the optimal solution/policy.[34,35]

To formulate the ensemble selection problem as an RL task, it is necessary to define the following key components of the model. The agent is our proposed ensemble selection algorithm. We define the environment as a deterministic one, in the sense that taking the same action in the same state on two different occasions cannot result in different next states and/or different reinforcement values. More specifically, the environment in which our agent operates consists of all possible subsets of the n base predictors, each serving as a possible ensemble, thus consisting of 2^n states. The environment includes the empty set, which is considered the start state in our implementation. An example environment generated by five base predictors is shown in Fig. 1. It can be viewed as a lattice, and the arrows represent the actions the agent is allowed to take at each state.

The agent investigates the environment by moving from one state (one ensemble) to another in search of better rewards. The reward $R(s_t, a_t, s_{t+1})$ received for the transition from the state s_t to the new state s_{t+1} by executing the action a_t is calculated based on the predictive performance of the ensemble(s) involved in the action. In our experiments, performance is assessed on a validation set separate from the training set, as explained in Section 4. The agent begins its learning at the start state (which corresponds to the empty set) without any prior knowledge about the subsequent states or any of the rewards. Then, the agent moves downward through the lattice from one state to another, until it reaches the final state, *i.e.*, the full ensemble

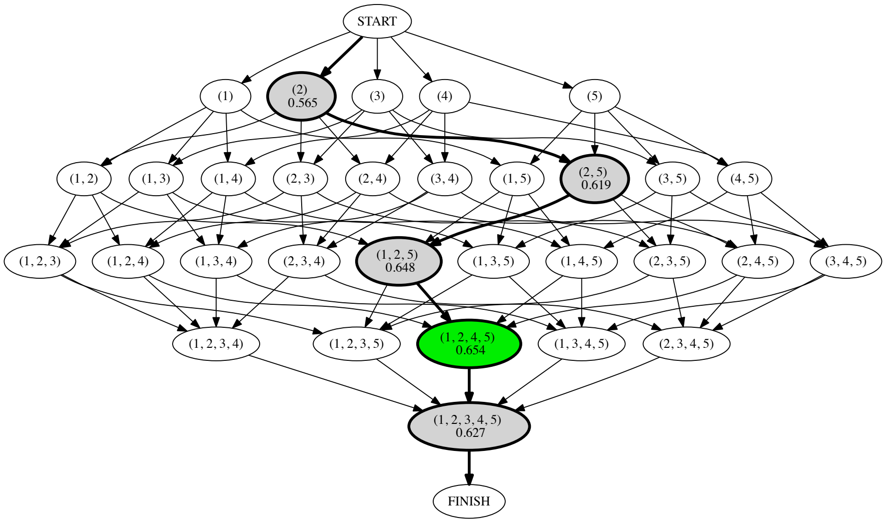


Fig. 1. Example of an environment constructed from five base predictors for the *P. pacificus* splice site dataset. The numbers in the nodes represent the predictive performance of the corresponding ensemble in terms of F-max on the validation set. The path highlighted in grey represents the policy found at the end of the learning process using the RL_greedy strategy ($\epsilon = 0.01$ and ten consecutive episodes yield the same ensemble). The state with the highest reward (*i.e.*, the highest predictive performance) along the path is (1, 2, 4, 5) (green), which is returned by the RL_greedy as the resulting ensemble for the illustrated example. The nodes (2, 4, 5) and (1, 2) represent the ensembles returned by the RL_pessimistic and RL_backtrack strategies respectively. Figure created with Graphviz.[36]

containing all base predictors. Each action is equivalent to adding exactly one other base predictor to the ensemble. In our model, the agent cannot "jump" from a state of k base predictors to a state of $k + 1$ base predictors that is not directly connected in the lattice, such as from (2, 3) to (1, 3, 4). This is necessary to control the complexity of the state-action space. A "traversal" of the lattice from the start state to the finish state is referred to as a "learning episode". We propose Q-learning-based algorithms that employ three different search strategies formulated as different ways of constructing learning episodes and computing rewards. These strategies essentially represent various models of the environment and the interactions between the agent and the environment. The goal, *i.e.*, traversing the lattice and ultimately reaching the full ensemble state, remains the same. These strategies are explained below.

3.1. *Greedy strategy (RL_greedy)*

Our first strategy emulates a "greedy" agent, whose goal is to reach the full ensemble quickly. As a consequence of this plan, the agent rapidly accumulates new base predictors and greedily constructs the ensemble. With each step taken from state s_t to s_{t+1} in the environment, the agent receives state s_{t+1}'s performance as a reward, *i.e.*, $R(s_t, a_t, s_{t+1}) = f(s_{t+1})$, and updates its Q-table as per Equation 1. Each learning episode has a fixed number of steps determined by the depth of the lattice. Fig. 1 shows an example of the path, as well as the resultant ensemble, that RL_greedy finds from a sample lattice constructed from five base predictors.

3.2. *Pessimistic strategy (RL_pessimistic)*

The second strategy resembles a "pessimistic" agent that resets itself to the start state as soon as a less than desirable state is visited, without finishing the traversal of the lattice, and starts a new learning episode. Thus, the lengths of the learning episodes may vary, depending on how soon the agent encounters a "depreciated" state. This strategy is based on the hypothesis that such a depreciated state (negative reward) might indicate a path of overfit and/or under-performing ensembles. For this strategy, the reward is calculated as the difference in performance from s_t to s_{t+1}, *i.e.*, $R(s_t, a_t, s_{t+1}) = f(s_{t+1}) - f(s_t)$.

3.3. *Backtrack strategy (RL_backtrack)*

The last strategy, RL_backtrack, uses the same reward function as RL_greedy. However, unlike the latter strategy, which aims to reach the full ensemble state quickly, the agent "backtracks" (goes back) to the immediately previous state as soon as a decrease in performance is encountered, and resumes its learning from there. A feature of this strategy is that the learning episodes can have a variable numbers of steps, as the agent might wander inside the lattice until it finds an acceptable path ending in the full ensemble.

For all the strategies, the policy derived from the learned action-value function yields a sequence/path of increasingly larger ensembles. We choose the ensemble on this path with the highest performance as the selected ensemble to be evaluated on the test set.

4. Experimental Setup

In all our RL-based experiments, the parameters of the Q-learning algorithm described in Section 2.4, namely α and γ are set to 0.1 and 0.9 respectively, which are commonly used values.[26] The exploration/exploitation trade-off is controlled by the ϵ probability discussed in Section 2.3. A higher probability indicates more exploration. In our work, we experiment with $\epsilon \in \{0.01, 0.1, 0.25, 0.5\}$ for the PF datasets and with $\epsilon \in \{0.01, 0.1, 0.25\}$ for the larger SS ones. The iterative nature of Q-learning requires the initialization of its parameters (values in the Q-table). We initialize the Q-table as a *zero* matrix, and update the values as states and rewards are observed by the agent, as guided by the search strategies defined above.

Although convergence of the Q-learning algorithm and the final policy "optimality" are theoretically guaranteed,[34,35] and achieved when the agent visits all of the environment's states infinitely often, we adopt more practical termination criteria for our search strategies in our experiments. For RL_greedy, the stopping point is reached when the policy induced by the agent produces the same result (*i.e.*, the same ensemble with the highest performance within the currently selected policy) for ten consecutive episodes. In contrast, RL_pessimistic and RL_backtrack are susceptible to longer running times, and even oscillations within the lattice and subsequently of the Q-values learned. For this reason, we set the termination criterion for these strategies as when the agent has taken a fixed number of steps in the environment, specifically 0.5 million. Note that these practical assumptions make it difficult for us to assess the theoretical optimality of our RL algorithms and their results.

In order to assess the relative performance of our RL-based ensemble selection strategies, we also employ several baselines. First, we consider the best base predictor (BP) of each initial set of base classifiers, which is the classifier with the highest classification prediction performance on the validation set. At the opposite end of the spectrum, the full ensemble (FE), consisting of all initial base classifiers, will be the largest ensemble, and our second baseline. Finally, the third baseline is the ensemble produced by CES, implemented as described in Section 2.2.

The general workflow used for the experimental evaluation of the above approaches is shown in Fig. 2. We use 5-fold cross validation to estimate the performance of all the models. All base predictors are learned on the training set (60% of the original data described in Table 1), which is balanced using undersampling of the majority class. The validation set (20% of the data) is used for calculating the rewards of the nodes in the RL environment, and to estimate the performance of the candidate base predictors in the CES approach. The test set (comprising the remaining 20% of the data) is used to assess the overall performance of all studied algorithms. An experiment for an algorithm being tested consists of the collection of these performance scores over all five rounds of this cross-validation.

All performance evaluation, whether internal (on the validation set) or external (on the test set) is conducted using F-max, which is the maximum value of the F-measure across all the values of precision and recall at many thresholds applied to the prediction scores generated by the base classifiers and the resultant ensembles. F-max is appropriate given the highly skewed class distributions of the datasets used in our study, and has been shown to be reliable for performance evaluation in a recent large-scale assessment of protein function prediction.[2] Other metrics that are sensitive to unbalanced problems, *e.g.*, area under the Precision-Recall curve, can also be used.

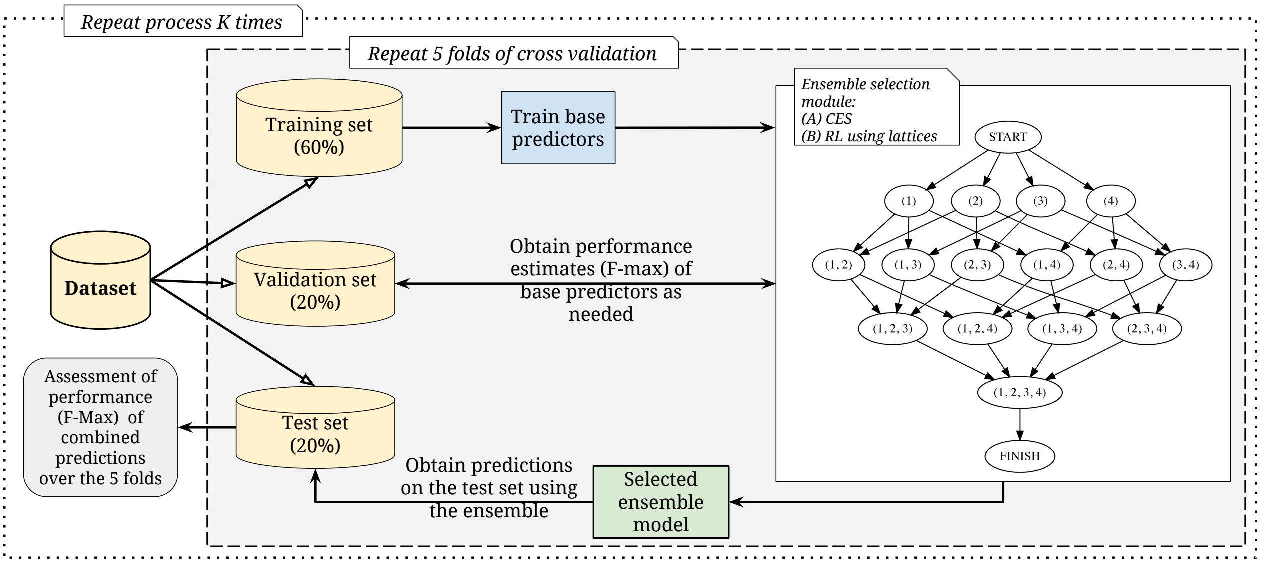


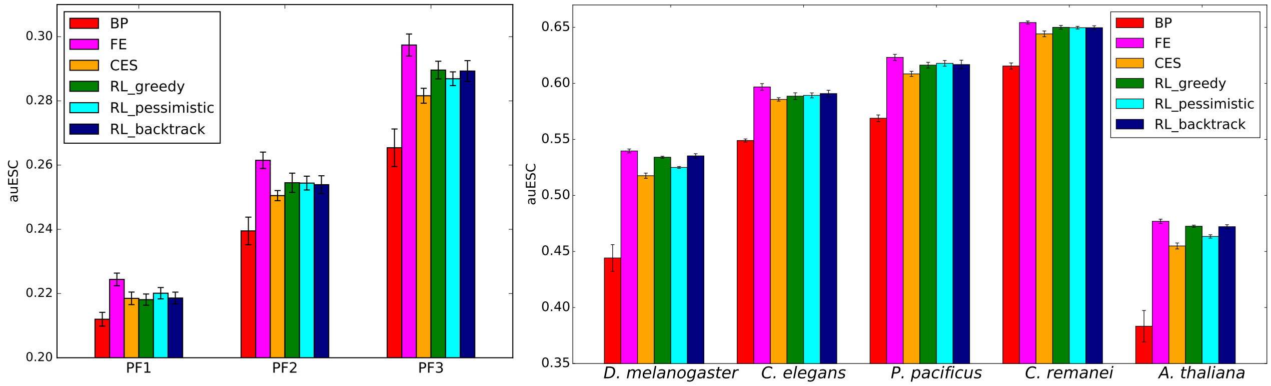
Fig. 2. Visual description of the workflow used to in our experiments. K = 10 for the PF datasets and K = 5 for the SS datasets.

The ensembles selected by the various approaches we tested (BP, RL, CES) are created by combining the probabilities produced by the constituent base predictors using a weighted average. Here, the importance (weight) of each base predictor is proportional to its predictive performance (measured in terms of the F-max score) on the validation set for all the approaches. We also considered options such as unweighted mean and median but observed that the performance of the ensembles was suffering because of the worst performing individual base predictors among them. In order to efficiently obtain the reward of each state or the performance of the ensemble being considered, we aggregate the predictions using a cumulative moving average.

We train 18 diverse base classification algorithms from Weka,[37] including Naïve Bayes, Multilayer Perceptron, SVM with a polynomial kernel, AdaBoost, Logistic Regression, and Random Forest. Each training set is resampled – to balance the classes – with replacement 10 times, resulting in 180 base predictors. Each ensemble selection algorithm is presented with a pool of base predictors (classification models). We start all our experiments with ten base predictors and increase this set gradually, with steps of ten randomly selected base predictors for each experiment, until we reach the entire set of 180 base predictors. This setup is designed to address the question of how the ensemble selection methods behave with an increasingly larger set of initial base predictors to select from. The performance of the ensembles resulting from of all these methods was evaluated across all these sizes, resulting in curves depicting the dependence of F-max on the number of initial base classifiers. To account for variation, each set of experiments was repeated ten times for the protein function datasets and five times for the splice site datasets (due to time constraints and the much larger size of the SS datasets). We used the area under these curves, denoted auESC (area under Ensemble Selection Curve), as a global evaluation measure for the various algorithms in our study, as it provides a global assessment of ensemble performance across a variety of base predictors. The area is normalized by its maximum possible value, which is the total number of base predictors (the maximum possible value of F-max is one); thus, the maximum value of auESC is one. However, this metric does not follow the same characteristics as auROC, such as random predictors/ensembles producing a score of 0.5. Therefore, auESC is mostly intended for comparative analyses between algorithms running on the same datasets (as done in our experiments), rather than for assessing the absolute performance of these algorithms.

5. Results

In this section, we will investigate various dimensions of classification performance and parsimony of ensembles constructed for the protein function (PF) and splice site (SS) prediction problems.



(a) PF datasets

(b) SS datasets

Fig. 3. Overall performance (as the auESC score described in Section 4) of all tested algorithms, evaluated across 18 sizes of initial base predictors (from ten to 180, in steps of ten), for (a) PF and (b) SS datasets. The standard errors are calculated over ten repetitions of each experiment for the PF datasets and five repetitions for the SS datasets. The first three bars of each dataset belong to the baselines considered, namely BP (best single base predictor), FE (full ensemble), and the CES algorithm. The last three bars of each dataset represent the RL-based approaches, namely RL_greedy (ten consecutive episodes yielding the same ensemble), RL_pessimistic, and RL_backtrack (both with 0.5 million training steps). For all RL-based approaches, the exploration/exploitation probability $\epsilon = 0.01$.

5.1. *Overall performance of ensemble selection methods*

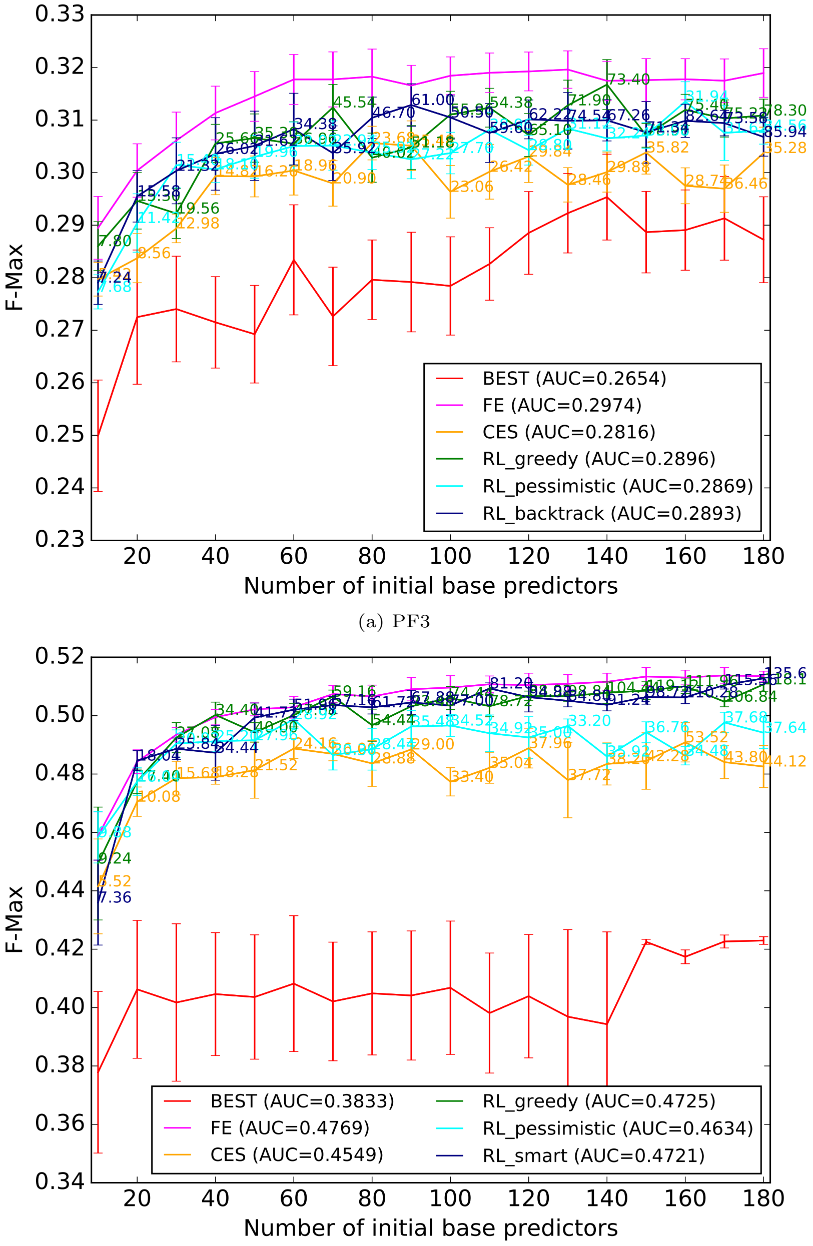
We first compared the overall classification performance of all the tested ensemble selection algorithms on all the datasets considered. These results are presented in the form of auESC scores in Fig. 3. We also tested the statistical significance of the comparisons of these algorithms in terms of these scores using the Friedman-Nemenyi tests.[38] Several trends can be observed from these figures.

- The best single base predictor (BP) is consistently outperformed by the full ensemble (FE $p = 6.805 \times 10^{-20}$ for the PF datasets and $p = 2.59 \times 10^{-15}$ for the SS datasets) and the RL-based approaches ($p_{RL\_greedy} = 3.29 \times 10^{-4}$, $p_{RL\_pessimistic} = 1.33 \times 10^{-6}$, $p_{RL\_backtrack} = 7.87 \times 10^{-5}$ for the PF datasets and $p_{RL\_greedy} = 5.27 \times 10^{-9}$, $p_{RL\_pessimistic} = 1.105 \times 10^{-5}$, $p_{RL\_backtrack} = 2.78 \times 10^{-8}$ for the SS datasets), thus validating the benefit of ensembles for these problems. In contrast, CES does not show statistically significant improvement over BP ($p > 0.05$ for both types of datasets).
- The biggest ensemble consisting of all the base predictors together (FE) achieves the highest performance across all tested datasets ($p < 0.05$ for pairwise comparisons with all other approaches.)
- For the smaller PF datasets, CES and the RL-based approaches are comparable ($p > 0.05$ for all pairwise comparisons). For the much larger SS datasets, the RL-based RL_greedy and RL_backtrack approaches perform significantly better than CES ($p_{RL\_greedy} = 0.01$, $p_{RL\_backtrack} = 0.025$), while RL_pessimistic does not ($p > 0.05$).
- Among the RL-based approaches, RL_greedy produces the best overall performance in terms of auESC, but the performance of all these approaches is statistically comparable ($p > 0.05$ for all pairwise comparisons for both PF and SS datasets.)

5.2. *Detailed examination of ensemble characteristics over various ensemble sizes*

The analysis based on auESC values shows the overall comparison of the classification performance of the approaches. However, it is critical to examine in detail the dynamics of these ensembles as the number of initial base predictors increases, and as a result, the sizes of the ensembles learnt. Due to lack of space, we only show two representative sets of curves in Fig. 4, one for the PF3 dataset and another for the splice site dataset from *A. thaliana*, showing the variation of ensemble F-max as the number of base predictors increases. These datasets were selected as representatives, as the ensembles showed the most benefit for them over individual base predictors (Fig. 3).

Fig. 4 shows that the performance of several ensemble approaches improves as the number of initial base



(a) PF3

(b) Splice site dataset from *A. thaliana*

Fig. 4. Performance in terms of F-max of all ensembles as the number of initial base predictors increases from ten to 180 in steps of ten. Each curve corresponds to averages over ten repetitions of each experiment for the (a) PF3 dataset and five repetitions for the (b) *A. thaliana* splice site dataset, along with standard error bars. The curves are annotated with average sizes of the ensembles learnt by the various algorithms at the points shown.

predictors increases, indicating that they are better able to utilize the information learnt. In particular, the curves for the RL approaches track much closer to the FE ones as compared to the CES and BP curves. These observations are stronger for the larger SS datasets (Fig. 4(b)) compared to the smaller PF ones (Fig. 4(a)). Furthermore, the substantial error bars on the CES curves indicate the method's ad-hoc nature. These results show that our RL-based approaches are better able to utilize the information in larger datasets than CES to produce more accurate and stable predictions.

As emphasized earlier, a desirable characteristic of any ensemble selection method is the ability to discover/construct a parsimonious ensemble. To assess this ability of the ensemble selection algorithms we tested,

Table 2. Statistics for the curves shown for PF3 in Fig. 4(a). Column auESC shows the overall performance of the algorithm over all sizes of initial base predictors sets. The ratios of the size and performance of the produced ensembles to those of the best performing approach FE are shown at representative initial base predictors set sizes of 60, 120, and 180.

	auESC	size_ratio@60	size_ratio@120	size_ratio@180	perf_ratio@60	perf_ratio@120	perf_ratio@180
BP	0.2654	0.0167	0.0083	0.0056	0.8919	0.9037	0.9005
FE	0.2974	1	1	1	1	1	1
CES	0.2816	0.31	0.24	0.19	0.9454	0.9493	0.9523
RL_greedy	0.2896	0.62	0.54	0.43	0.9615	0.9621	0.9746
RL_pessimistic	0.2869	0.37	0.24	0.19	0.9601	0.9531	0.9656
RL_backtrack	0.2893	0.57	0.52	0.48	0.9702	0.9714	0.9619

Table 3. Statistics for the curves shown for *A. thaliana* in Fig. 4(b). Column auESC shows the overall performance of the algorithm over all sizes of initial base predictors sets. The ratios of the size and performance of the produced ensembles to those of the best performing approach FE are shown at representative initial base predictors set sizes of 60, 120, and 180.

	auESC	size_ratio@60	size_ratio@120	size_ratio@180	perf_ratio@60	perf_ratio@120	perf_ratio@180
BP	0.3833	0.0167	0.0083	0.0056	0.8118	0.7912	0.8237
FE	0.4769	1	1	1	1	1	1
CES	0.4549	0.40	0.31	0.24	0.9710	0.9577	0.9379
RL_greedy	0.4725	0.50	0.50	0.51	0.9946	0.9927	0.9945
RL_pessimistic	0.4634	0.48	0.29	0.21	0.9919	0.9649	0.9623
RL_backtrack	0.4721	0.87	0.79	0.75	0.9983	0.9919	0.9985

we show in Tables 2 and 3 important statistics of the curves shown in Fig. 4. Specifically, we compute ratios of the selected ensembles' performance and sizes to those of the best performing and largest ensemble, FE. For brevity, we only show statistics at the representative initial base predictors set sizes of 60, 120, and 180.

From the results shown in Tables 2 and 3, it can be seen that CES discovers the smallest ensembles, but these appear not to have enough predictive information, resulting in lower classification performance than FE, especially in the case of the *A. thaliana* SS dataset. The RL-based approaches produce relatively larger ensembles due to their ability to search a much larger portion of the space of ensembles. Due to this same ability, they are able to select ensembles that produce classification performance nearly identical to FE, as shown by the performance ratios. Furthermore, the parsimony ability of these ensembles is enhanced as the number of initial base predictors increases, without significant variation in classification performance. This again validates the parsimonious yet accurate characteristic of the RL ensembles. Finally, our "pessimistic" strategy achieves the best parsimony-performance balance, possibly because of earlier resets of the learning episodes that force the agent to evaluate the upper levels of the lattice, where the smaller ensembles reside. We recommend analyzing summary statistics like the above to identify the ensembles representing the best parsimony-performance balance obtained from various ensemble selection methods. From another perspective, these ensembles might be identified as the point(s) at which the curve(s) like the ones in Fig. 4 start plateauing.

5.3. *Dependence of ensemble selection algorithms' behavior on parameters*

The exploration probability ϵ is critical to the RL-based approaches as it controls the exploration/exploitation management, and consequently how much of the ensemble space is visited. To assess the effect of this parameter on the RL ensembles, we evaluated all three search strategies by executing them with $\epsilon \in \{0.01, 0.1, 0.25, 0.5\}$ for the PF datasets, and with $\epsilon \in \{0.01, 0.1, 0.25\}$ for the SS datasets. We then conducted ANOVA with the F-test to assess the effect of the ϵ values on both ensemble classification and size over the whole performance curves of the type shown in Fig. 4.

For the RL ensemble results from the PF3 dataset (Fig. 4(a)), both in terms of classification performance ($p = 0.26$) and ensemble size ($p = 0.75$), the RL_greedy approach produces similar results for different ϵ values. The RL_pessimistic and RL_backtrack strategies, however, produce statistically variable results with different epsilons, both in terms of classification performance ($p_{RL\_pessimistic} = 7.3 \times 10^{-10}$, $p_{RL\_backtrack} = 0.01$) and ensemble size ($p_{RL\_pessimistic} = 2.7 \times 10^{-13}$, $p_{RL\_backtrack} = 3.85 \times 10^{-4}$). The same trends are observed for PF1 and PF2. Further analysis of RL_pessimistic and RL_backtrack shows that as ϵ increases, the resultant ensemble performance drops, suggesting that too much exploration of the ensemble space may lead to overfitting.

The same analysis of the SS datasets also yielded similar results, with the exception that RL_backtrack did not show a dependence on the value of ϵ for any of the datasets, both in terms of classification performance (*e.g.*, $p = 0.87$ for *A. thaliana*) and ensemble size (*e.g.*, $p = 0.47$ for *A. thaliana*). Given the above observations, as well as the speed of execution, we show results only for $\epsilon = 0.01$ in the previous subsections.

Finally, as mentioned in Section 2.2, CES was run using the recommended values for its parameters.[21,22] To test the impact of these parameters, which are generally difficult to set a priori, we vary their values as the pool size of candidate base predictors $m \in \{N/2, N\}$ and the maximum ensemble size $\in \{N/4, N/2, 3N/4, N\}$ (N = total number of base predictors), and conducted a similar ANOVA analysis as above. The pool size did not have a significant impact on the performance, or the size of the ensembles ($p > 0.05$ for both PF3 and *A. thaliana*). However, the performance is significantly sensitive to the parameter controlling the maximum ensemble size ($p = 3.8 \times 10^{-3}$ for PF3 and $p = 2.8 \times 10^{-4}$ for *A. thaliana*), which also influences the size of the resulting ensemble slightly ($p = 0.08$ for PF3 and $p = 0.01$ for *A. thaliana*). Considering the above results with these, it can be inferred that CES is indeed more sensitive to its key parameters than the RL-based approaches, especially the best performing RL_greedy.

6. Discussion

Ensemble selection offers a powerful approach to addressing biomedical prediction problems, as well as gaining novel knowledge by the analysis of the selected ensembles. This paper presents a framework for selecting ensembles of classifiers using elements of reinforcement learning (RL), which offers a systematic and mathematically sound methodology for exploring the many possible combinations of base predictors that can be selected into an ensemble. This is in contrast to existing ensemble selection algorithms like CES, which often make ad-hoc decisions during ensemble learning and thus cannot offer performance guarantees. Several RL-based methods were implemented in our framework to search the space of possible ensembles as exhaustively as needed.

We tested our methods on two computational genomics problems, namely protein function prediction and splice site detection. We observed that our proposed RL-based methods are able to capture predictive performance close to the full ensembles with a much smaller number of base predictors. This observation is especially strong for the larger splice site datasets, an outcome worth investigating for other biomedical problems, such as cancer phenotype prediction. We also observed that many of the RL parameters, such as the exploration/exploitation probability (ϵ), did not have a significant impact on the downstream performance or sizes of the selected ensembles. It is necessary to examine the intrinsics of RL approaches, such as the effect of ϵ, on a variety of datasets to assess their strengths and weaknesses more comprehensively. Even more fundamentally, the RL approaches we tested can be reformulated, such as by revising the constituent reward functions, to yield better and/or more insightful ensembles. A particular avenue of interest here is to analyze the diversity of the base predictors selected into ensembles by the RL approaches in greater depth, and how that contributes to ensemble performance. Finally, there is also a need to investigate how the performance of our RL algorithms relates to the optimization capabilities and performance guarantees offered by Q-learning.

The RL algorithms studied allow for a larger portion of the space of possible ensembles to be examined more systematically, but this also causes computational requirements to grow substantially, especially as ϵ, the exploration probability, and the number of initial base predictors, grow. For instance, with our basic implementation of these algorithms, which are available as open-source software from `https://github.com/GauravPandeyLab/`, the time (on a 2.3 GHz processor) for one *A. thaliana* RL_greedy experiment varies from approximately one second with ten base predictors to approximately 3 minutes with 180 base predictors for $\epsilon = 0.01$. The same kind of executions took approximately 20 minutes on average for 180 base predictors with $\epsilon = 0.5$. For RL_pessimistic and RL_backtrack, the recorded times for an experiment with 180 base predictors and $\epsilon = 0.01$ are substantially higher, approximately 5.5 and 9.25 hours respectively, due to their more extensive exploration and resets in the search process. The memory requirements also vary similarly with the number of initial base predictors and value of ϵ. To make such executions more computationally feasible, especially for larger datasets, there is a need for developing more parallelized/optimized implementations of these algorithms. We will release such implementations in the future, and invite the community to participate in this effort.

To conclude, our overall effort was towards constructing accurate yet parsimonious (smaller) ensembles, which may in turn be more interpretable, for difficult problems. We acknowledge that interpretation can be a challenging and often subjective task, depending on the target problem, which is why it was out of the scope of our paper, and needs a deeper investigation. Although we didn't explicitly consider this, the interpretability of the base predictors itself would have a major impact on the interpretability of their resultant ensemble. Indeed, this interpretability criterion can be explicitly considered during the ensemble selection/search process to address this challenge more directly.

7. Acknowledgements

This work was partially supported by NIH grant # R01-GM114434 and an IBM faculty award to GP. We thank the Icahn Institute for Genomics and Multiscale Biology and the Minerva supercomputing team for their financial and technical support. We also thank Om P. Pandey and Gustavo Stolovitzky for their technical advice, and the anonymous reviewers for their comments and suggestions.

References

1. G. Pandey, V. Kumar and M. Steinbach, Computational Approaches for Protein Function Prediction: A Survey, Tech. Rep. 06-028, University of Minnesota (2006).
2. P. Radivojac, W. T. Clark, T. R. Oron et al., Nature methods **10**, 221 (2013).
3. M. Kuhn, M. Campillos, P. González, L. J. Jensen and P. Bork, FEBS letters **582**, 1283 (2008).
4. G. Schweikert, G. Rätsch, C. Widmer and B. Schölkopf, Adv. in Neural Info. Processing Systems **22**, 1433 (2009).
5. L. Rokach, Artificial Intelligence Review **33**, 1 (2009).
6. G. Seni and J. F. Elder, Synthesis Lectures on Data Mining and Knowledge Discovery **2**, 1 (2010).
7. P. Yang, Y. H. Yang, B. B. Zhou and A. Y. Zomaya, Current Bioinformatics **5**, 296 (2010).
8. A. Altmann, M. Rosen-Zvi, M. Prosperi et al., PLoS ONE **3**, p. e3470 (2008).
9. A. Khan, A. Majid and T.-S. Choi, Amino Acids **38**, 347 (2010).
10. G. Pandey, B. Zhang, A. N. Chang et al., PLoS Computational Biology **6**, p. e1000928 (2010).
11. G. Yu, H. Rangwala, C. Domeniconi, G. Zhang and Z. Yu, Trans. on Comp. Biol. and Bioinfo. **10**, 1045 (2013).
12. Y. Guan, C. Myers, D. Hess et al., Genome Biology **9**, p. S3 (2008).
13. M. M. Ward, S. Pajevic, J. Dreyfuss and J. D. Malley, Arthritis Care & Research **55**, 74 (2006).
14. K. Tumer and J. Ghosh, Connection Science **8**, 385 (1996).
15. L. I. Kuncheva and C. J. Whitaker, Machine Learning **51**, 181 (2003).
16. T. G. Dietterich, Machine Learning **40**, 139 (2000).
17. R. E. Schapire and Y. Freund, Boosting: Foundations and Algorithms (MIT Press, 2012).
18. L. Breiman, Machine learning **45**, 5 (2001).
19. C. J. Merz, Machine Learning **36**, 33 (1999).
20. D. H. Wolpert, Neural Networks **5**, 241 (1992).
21. R. Caruana, A. Niculescu-Mizil, G. Crew and A. Ksikes, Intl. Conference on Machine Learning **21**, p. 18 (2004).
22. R. Caruana, A. Munson and A. Niculescu-Mizil, International Conference on Data Mining **6**, 828 (2006).
23. S. Whalen and G. Pandey, International Conference on Data Mining **13**, 807 (2013).
24. S. Whalen, O. P. Pandey and G. Pandey, Methods **93**, 92 (2016).
25. A. Niculescu-Mizil, C. Perlich, G. Swirszcz, V. Sindhwani and Y. Liu, J. of Mach. Learn. Research **7**, 23 (2009).
26. R. S. Sutton and A. G. Barto, Intro to Reinforcement Learning, 1st edn. (MIT Press, Cambridge, MA, USA, 1998).
27. T. R. Hughes, M. J. Marton, A. R. Jones et al., Cell **102**, 109 (2000).
28. C. L. Myers, D. R. Barrett, M. A. Hibbs, C. Huttenhower and O. G. Troyanskaya, BMC Genomics **7** (2006).
29. M. B. Shapiro and P. Senapathy, Nucleic acids research **15**, 7155 (1987).
30. G. Rätsch, S. Sonnenburg, J. Srinivasan et al., PLoS Comput Biol **3**, p. e20 (2007).
31. J. Kober, J. A. Bagnell and J. Peters, International Journal of Robotics Research **32**, 1238 (2013).
32. M. L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming (1994).
33. C. J. C. H. Watkins, Learning from delayed rewards, PhD thesis, University of Cambridge England1989.
34. C. J. C. H. Watkins and P. Dayan, Machine Learning **8**, 279 (May 1992).
35. D. P. Bertsekas and J. N. Tsitsiklis, Neuro-Dynamic Programming, 1st edn. (Athena Scientific, 1996).
36. E. R. Gansner and S. C. North, SOFTWARE - PRACTICE AND EXPERIENCE **30**, 1203 (2000).
37. M. Hall, E. Frank, G. Holmes et al., ACM SIGKDD Explorations Newsletter **11**, 10 (2009).
38. J. Demšar, Journal of Machine Learning Research **7**, 1 (2006).

METHODS FOR CLUSTERING TIME SERIES DATA ACQUIRED FROM MOBILE HEALTH APPS

NICOLE TIGNOR[1], PEI WANG[1], NICHOLAS GENES[1,2], LINDA ROGERS[3], STEVEN G. HERSHMAN[4], ERICK R. SCOTT[1], MICOL ZWEIG[1], YU-FENG YVONNE CHAN[1,2], ERIC E. SCHADT[1]

[1]*Department of Genetics and Genomic Sciences*
Icahn School of Medicine at Mount Sinai, New York, NY, USA

[2]*Department of Emergency Medicine*
Icahn School of Medicine at Mount Sinai, New York, NY, USA

[3]*Department of Medicine, Pulmonary, Critical Care and Sleep Medicine*
Icahn School of Medicine at Mount Sinai, New York, NY, USA

[4]*LifeMap Solutions, Inc, New York, NY, USA*

Email: pei.wang@mssm.edu, eric.schadt@mssm.edu

In our recent Asthma Mobile Health Study (AMHS), thousands of asthma patients across the country contributed medical data through the iPhone Asthma Health App on a daily basis for an extended period of time. The collected data included daily self-reported asthma symptoms, symptom triggers, and real time geographic location information. The AMHS is just one of many studies occurring in the context of now many thousands of mobile health apps aimed at improving wellness and better managing chronic disease conditions, leveraging the passive and active collection of data from mobile, handheld smart devices. The ability to identify patient groups or patterns of symptoms that might predict adverse outcomes such as asthma exacerbations or hospitalizations from these types of large, prospectively collected data sets, would be of significant general interest. However, conventional clustering methods cannot be applied to these types of longitudinally collected data, especially survey data actively collected from app users, given heterogeneous patterns of missing values due to: 1) varying survey response rates among different users, 2) varying survey response rates over time of each user, and 3) non-overlapping periods of enrollment among different users. To handle such complicated missing data structure, we proposed a probability imputation model to infer missing data. We also employed a consensus clustering strategy in tandem with the multiple imputation procedure. Through simulation studies under a range of scenarios reflecting real data conditions, we identified favorable performance of the proposed method over other strategies that impute the missing value through low-rank matrix completion. When applying the proposed new method to study asthma triggers and symptoms collected as part of the AMHS, we identified several patient groups with distinct phenotype patterns. Further validation of the methods described in this paper might be used to identify clinically important patterns in large data sets with complicated missing data structure, improving the ability to use such data sets to identify at-risk populations for potential intervention.

1. Introduction

Handheld mobile devices such as the smartphone are increasingly being utilized by app developers to help users better manage their health and chronic disease conditions. These devices and the mobile health apps that run on them have the potential to provide critical, longitudinal components to an individual's health record. In fact companies such as Apple have greatly facilitated this through their HealthKit, ResearchKit, CareKit, and Homekit platforms, which enable acquisition of very high frequency data over long periods of time, thus providing far more detailed phenotypic user profiles than could ever be reasonably generated in a typical clinical or research setting.

Recently, benefiting from advances in mobile health technologies, we successfully conducted the Asthma Mobile Health Study using an iPhone app.[1] Asthma is a common, highly variable and heterogeneous disease, and it has therefore been difficult to characterize patient disease subtypes precisely enough to inform an optimal individualized treatment plan. Less than half of the 25 million people in the United States with asthma have optimal asthma control, significantly contributing to $56 billion in direct and indirect health care costs annually.[2-3] In order to improve outcomes and reduce costs on a population level, it will be important to acquire large data sets to develop individualized models capable of identifying patients at highest risk to better target resources and tailor therapies. Prior efforts at identifying subgroups of asthma patients have been made based on demographics, lung function tests, biopsy results and blood testing, response to therapy,[4] and recently, genetics.[5-6] Our Asthma Health App, however, for the first time, enables one to collect rich time series data on asthma patients' activities on a daily basis. This opens up the possibility to identify at-risk subgroups of patients based on high-resolution time-course symptom data. The ability to identify clinically relevant patterns of disease could potentially allow targeting of resources to at risk patients to improve disease control.

Participants in the Asthma Mobile Health Study (AMHS) were asked to complete daily surveys to record symptoms and presumed triggers for the duration of the study. Taking the *day symptom outcome* as an example, the collected data of one user is a vector of 0's and 1's indicating whether the user experienced any asthma symptom on each day (1 indicating yes and 0 no symptom experienced). Once collected, the day symptom outcome records of all users can be presented as a 0/1 matrix, which can be used to explore whether subgroups of asthma patients with distinct symptom patterns exist. However, one particular challenge with this type of survey results data is that they contain substantial missing values. While most users may respond to daily survey questions or choose to actively input data on their condition when appropriate, for any given subset of days for which data are being collected, the response rates will be highly varied among different users. Further, for formal studies such as AMHS, users enroll in the study on a rolling basis, such that many non-overlapping periods of enrollment among different users must be accounted for. Lastly, even for the same user, survey response rates often varied over time. Users may be more likely to respond on days when they experience disease symptoms, which further complicates analysis of the data.

A crucial step in handling missing data is to characterize the nature of missing-ness. If the probability of missing data does not depend on the missing values, the missing-data mechanism is referred to as missing-at-random; if so, the mechanism is referred to as not-missing-at-random or non-ignorable. When the proportion of missing values in a data set is large and the missing mechanism is not at random, it is not appropriate to ignore the missing mechanism and perform standard statistical analyses based on the observed values.[7-8] In our AMHS data, since the probability of a user responding to the survey on a particular day depends on the user's asthma symptom on that day, the missing mechanism is non-ignorable. Therefore, in this work, we propose a probability model to characterize the missing mechanism underlying such data and implement a consensus clustering algorithm incorporating multiple imputations. We compare our proposed method with other imputation strategies based on low rank matrix completion procedures.[9] Through extensive simulation studies, we demonstrate the advantage of the probability model based imputation under a range of scenarios reflecting the characteristics of our time series data. While our method is applied to AMHS study and simulated data, the approach can be applied to any time series data in which the missing data mechanism is non-ignorable.

2. Method

Our primary aim was to develop a method that would cluster users in AMHS based on their self-reported day symptom outcome time-series data to identify subgroups of app users with distinct symptom patterns. Given the substantial amount of missing data and that the missing data mechanism is non-ignorable, existing methods were not sufficient for this purpose.

2.1. *A probability based imputation model*

Denote the *day symptom outcome* data matrix as $X_{N\times T} = [\![x_{it}]\!]$, where $i = 1,\cdots,N,$ is the index of users and $t = 1,\cdots,T,$ is the index of days. Note, since asthma symptoms are often affected by environmental and seasonal changes, we align the profiles of different users according to actual dates instead of arbitrary days in the study. Each x_{it} takes on a value of 1 or 0, depending on whether the i^{th} user reported an asthma symptom on the t^{th} day or not, respectively; x_{it} is set to NA if the i^{th} user did not enroll in the study or did not respond to the daily survey on the t^{th} day.

We further introduce two binary data matrices: $S_{N\times T} = [\![s_{it}]\!]$ to indicate whether users responded to the AMHS survey on each day; and $D_{N\times T} = [\![d_{it}]\!]$ to represent the underlying complete day symptom outcome data. Given these matrices, the observed data $X_{N\times T}$ satisfies: $x_{it} = d_{it}$, if $s_{it} = 1$; and $x_{it} = NA$ if $s_{it} = 0$. If $D_{N\times T}$ was available, existing methods could be employed to cluster users based on this data matrix. However, since we only observe $X_{N\times T}$ and a substantial proportion of $X_{N\times T}$ is NA, we need to impute these missing values first before we can attempt clustering.

The key step for the imputation is to estimate the probability that a given user on a given day had a symptom event that should have been recorded, given the user did not respond to the survey on that day $P(d_{it} = 1|s_{it} = 0)$. In light of the 6-month milestone survey, which is administered to each app user 6 months after the enrollment date in our AMHS, 12% of users indicated that they were more likely to respond to the survey on days when they experienced asthma symptom(s). Given this, we assume that there exists an α_i (≥ 1) for each user, such that $P(s_{it} = 1|d_{it} = 1) = \alpha_i P(s_{it} = 1|d_{it} = 0) = \alpha_i r_{it}^0$, where $r_{it}^0 = P(s_{it} = 1|d_{it} = 0)$. We treat each α_i as a random variable, which takes the value of 1 with probability 0.88, and 2 with probability 0.12, in accordance with feedbacks from AMHS. The choice of 2 is based on the median level of possible range of α_i ($1 < \alpha_i < 3$) that ensures realistic scenarios given the observed distribution of user response rates. Sensitivity analysis on choices of α_i is shown in Section 3.

We further denote $\bar{p}_{it} = P(d_{it} = 1)$, $p_{it} = P(d_{it} = 1 | s_{it} = 1)$, and $r_{it} = P\ (s_{it} = 1)$. Thus we have that $r_{it} = P(s_{it} = 1 | d_{it} = 1) P(d_{it} = 1) + P(s_{it} = 1 | d_{it} = 0) P(d_{it} = 0) = \alpha_i r_{it}^0 \bar{p}_{it} + r_{it}^0 (1 - \bar{p}_{it}\ )$; $p_{it} = \frac{P(s_{it}=1|d_{it}=1)P(d_{it}=1)}{P(s_{it}=1|d_{it}=1)P(d_{it}=1)+P(s_{it}=1|d_{it}=0)P(d_{it}=0)} = \frac{\alpha_i \bar{p}_{it}}{\alpha_i \bar{p}_{it} + (1-\bar{p}_{it}\ )}$. it follows that

$$\bar{p}_{it} = \frac{p_{it}}{\alpha_i(1-p_{it})+p_{it}}, \quad \text{and} \quad r_{it}^0 = r_{it} \frac{\alpha_i(1-p_{it})+p_{it}}{\alpha_i}. \tag{1}$$

And we have that

$$P(d_{it} = 1 | s_{it} = 0) = \frac{P(s_{it}=0|d_{it}=1)P(d_{it}=1)}{P(d_{it}=0|s_{it}=1)P(s_{it}=1)+P(d_{it}=0|s_{it}=0)P(s_{it}=0)}$$

$$= \frac{(1-\alpha_i r_{it}^0)\bar{p}_{it}}{(1-\alpha_i r_{it}^0)\bar{p}_{it}+(1-r_{it}^0)(1-\bar{p}_{it})} = \frac{(1-\alpha_i r_{it}^0) p_{it}}{(1-\alpha_i r_{it}^0) p_{it}+\alpha_i(1-r_{it}^0)(1-p_{it})}. \tag{2}$$

We then propose to estimate p_{it} and r_{it}based on the observed data in a time window around the t^{th} day such that

$$\hat{p}_{it} = \frac{\sum_{|t' - t| < \delta} I(s_{it'} = 1,\ x_{it'} = 1)}{\sum_{|t' - t| < \delta} I(s_{it'} = 1)}, \text{and } \hat{r}_{it} = \frac{\sum_{|t' - t| < \delta} I(s_{it'} = 1)}{\sum_{|t' - t| < \delta} 1}, \tag{3}$$

where $I(\cdot)$ is the indicator function, and δ defines the size of the time window. If we plug equation (3) into equations (1) and (2), then we can obtain an estimate of $P(d_{it} = 1 | s_{it} = 0)$. In the simulation and real data analysis below, we set δ to be 30 days. This choice resulted from a tradeoff between the robustness to estimate empirical response/symptom rates and sensitivity to capture changes within a short time period.

2.2. *Multiple imputation and consensus clustering*

The probability model in section 2.1 provides a convenient framework for integrating the multiple-imputation procedure[8] and the consensus clustering procedure.[10] Specifically, in the b^{th} imputation run, we first simulate a vector of $\{\alpha_i^b\}_i$. Then to impute an unobserved d_{it}, we calculate $\widehat{P^b}(d_{it} = 1 | s_{it} = 0)$ based on α_i^b, and randomly sample a value from a Bernoulli distribution with success probability of $\widehat{P^b}(d_{it} = 1 | s_{it} = 0)$. We denote the final imputed complete matrix as ${D_{N \times T}}^b = [\![{d_{it}}^b]\!]$.

Naively, we could perform clustering analysis based on ${D_{N \times T}}^b$. However, when we compare the day symptom profiles of two users, it makes more sense to define distance based on their symptom frequencies over a time window instead of based on events on individual days. For example, suppose there are two users: one has symptoms on Monday, Wednesday and Friday in a given week, while the other has symptoms on Tuesday, Thursday and Saturday in the same week. If we considered the 0/1 vectors of daily symptom events of these two users for this week, they would be extremely different. However, if we consider the symptom frequency over the week, these two users actually show a similar pattern. Therefore, we propose to calculate the frequency profile of each user by performing a running average of the symptom profile:

$f_{it}^{b} = 1/(2h-1)\sum_{|t' -t|<h} s_{it'}^{b}$. Then, we can derive clusters of users by performing K-means clustering based on the frequency matrix $F_{N\times T}^{b} = [\![f_{it}^{b}]\!]$. We can record the clustering result with an adjacency matrix $(\left(A_{ij}{}^{b}\right))_{N\times N}$, where $A_{ij}{}^{b} = 1$ if the i^{th} user and the j^{th} user are assigned to the same cluster; and $A_{ij}{}^{b} = 0$ otherwise. We repeat the above imputation-cluster process B times. This gives us B adjacency matrices $\{(\left(A_{ij}{}^{b}\right))_{N\times N}\}_{b}$ corresponding to B sets of clustering results. Intuitively, a large value for A_{ij} suggests a high similarity between the i^{th} and j^{th} user. We can define an average adjacency matrix, $\overline{A_{ij}} = 1/B\sum_{b} A_{ij}{}^{b}$, over all adjacency matrices, and then perform the final cluster assignment via another round of K-mean clustering based on the $(\left(\overline{A_{ij}}\right))$ matrix. We refer to the above procedure as the probability based imputation with consensus clustering (PIC) method. For the special case of $h = 1$, clustering is performed on the imputed day symptom matrix $D_{N\times T}{}^{b}$. We refer to this special case as the PIC.s method.

One variation on the PIC method worth exploring is to first perform Principal Component Analysis (PCA) on the $D_{N\times T}{}^{b} = [\![d_{it}{}^{b}]\!]$ matrix, and then select the loading matrix of the leading L principle components to further perform the clustering analysis. We denote this variation of the PIC procedure as PIC.PC.

3. Simulation Studies

In this section, we investigate the performance of the proposed methods through simulation studies under a range of scenarios reflecting real data conditions.

3.1. *Methods to compare*

In addition to the three methods defined above, PIC, PIC.s and PIC.PC, we also consider performing the probability imputation without taking into account the non-random missing pattern (i.e. set $\alpha_i = 1$). We denote this strategy as "PIC($\alpha_i = 1$)". We also include a few low-rank (LR) matrix completion based approaches for comparison. LR matrix completion has been recently demonstrated to be extremely powerful in recovering large scale matrices[9]. Specifically, we employ the R package *softImpute*,[11] which uses convex relaxation techniques to provide regularized low-rank solutions for large-scale matrix completion problems. We considered three strategies to apply the LR matrix completion (referred to as "LR" in below): (1) we directly apply LR on the raw data matrix ($X_{N\times T}$); (2) for each user, we first imputed the missing data based on the probability model of PIC for days within his/her enrollment period, and then apply LR to impute the missing data on days outside the enrollment period; and (3) similar to (2) except that we further derive the frequency matrix following the imputations. Here, enrollment period of one user is defined as the period from the first to the last instance of non-missing observation based on the empirical day symptom data. In all three strategies, after data imputation, consensus clustering is performed in the same way as for PIC. We denote these three strategies as LR, PIC.S.LR, and PIC.LR, respectively.

3.2. *Simulation settings*

To mimic the data from AMHS, in our simulations (see section 4), we set N=334, T=136, and the total number of clusters to be 3. In addition, we assumed 3 roughly equal-sized clusters (n_1=111,

n_2=111, and n_3=112), so the accuracy of clustering result could be more intuitively assessed. We then generated multiple sets of frequency curves representing a variety of hypothetical symptom frequency profiles (i.e. $\{P\ (d_{it}=1)\}_t$) (see Fig. 1). We assume the samples belonging to the same cluster share the same underlying symptom frequency profile. To generate time-series data for each sample, we simulated symptom events of the t^{th} day by Bernoulli sampling of 0/1 based on the t^{th} point of the corresponding frequency curve. To simulate non-overlapping enrollment periods, we sampled from the empirically observed enrollment period distribution from the AMHS data.

To further generate non-ignorable missing-ness, we used information from the milestone survey results in AMHS. In this survey, users are asked to provide their reasons for not responding to the daily survey during the study period. Based on users who provided milestone survey responses before April 4, 2016, 12% indicated that they tended to skip the daily surveys on days in which they had no symptoms. Thus, in the simulated data we sampled from a Bernoulli distributed random variable I to identify whether a user was among those whose response depended on symptom state, with p (I = 1) = 0.12. For samples assigned to I = 1, we introduced a parameter Δ to modify the rate of missing data depending on symptom state such that $(s_{it}=1|d_{it}=1)=P(s_{it}=1)+\Delta$ and $P(s_{it}=1|d_{it}=0)=P(s_{it}=1)-\Delta$. For other samples assigned to I = 0, we imposed uniformity over time such that $P(s_{it}=1|d_{it}=1)=P(s_{it}=1|d_{it}=0)=P(s_{it}=1)$. For each user, $r_{it}=P(s_{it}=1)$ was set to be a constant r_i, which is either a pre-determined value or is sampled from an empirical distribution of missing rates calculated from the AMHS data. We then used $P(s_{it}=1|d_{it}=1)$ and $P(s_{it}=1|d_{it}=0)$ to generate missing data within the enrollment period.

We considered various simulation settings to evaluate how the performances of the different methods were affected by various factors including: (1) the shapes of the frequency profiles, (2) the overall missing percentages, (3) the severity level of the non-random missing, (4) alternative scenarios for setting α_i, and (5) we evaluated the power to detect association between a generic simulated covariate and the inferred cluster assignments derived from the application of each method on simulated data. In the following, we varied one factor at a time, where unless specified, the default setting is to use the frequency profile set labeled b in Figure 1, r_i = 0.4 for all users, Δ = 0.3, and α_i is 1 with probability 0.88 or 2 with probability 0.12. For all settings, the window size h used to derive the frequency profiles is simply set to be a fix value of 15, as we observed that the performance of all strategies are not sensitive to the different choices of h (data not shown).

1. We considered 8 different sets of symptom frequency profiles as illustrated in Figure 1.
2. We considered 4 different ways for setting r_i, where for (1)-(3), r_i = 0.2, 0.4, or 0.6; and for (4) r_i is sampled from an empirical distribution of missing rates calculated from the AMHS data.
3. We varied the value of Δ, where Δ = 0.1, 0.3, or 0.5.
4. We considered 3 alternative scenarios for setting α_i, where for (a1)-(a3): α_i is 1 with probability 0.88 and is 1.5, 2, or 2.5 with probability 0.12.
5. We simulated a binary covariate based on true cluster assignments, where the probability of taking a value of 1 was set to 10% across all clusters (p1), or was set, depending on cluster

assignment, to: (p2) 10%, 15% or 20%, or (p3) 10%, 20% or 30%. For these 3 scenarios, we evaluated the power to detect association between the simulated covariate and the predicted cluster assignments using a p-value cutoff of 0.05 based on Fisher's Exact Test.

3.3. *Simulation results*

For each simulation scenario, we applied each of the strategies in section 3.1 to derive predicted cluster assignments from simulated data sets. True and predicted cluster assignments were compared using the adjusted Rand index.[12] Based on the results from simulation Setting 1, strategies PIC and PICs perform well across a range of symptom profile scenarios (Fig. 1).

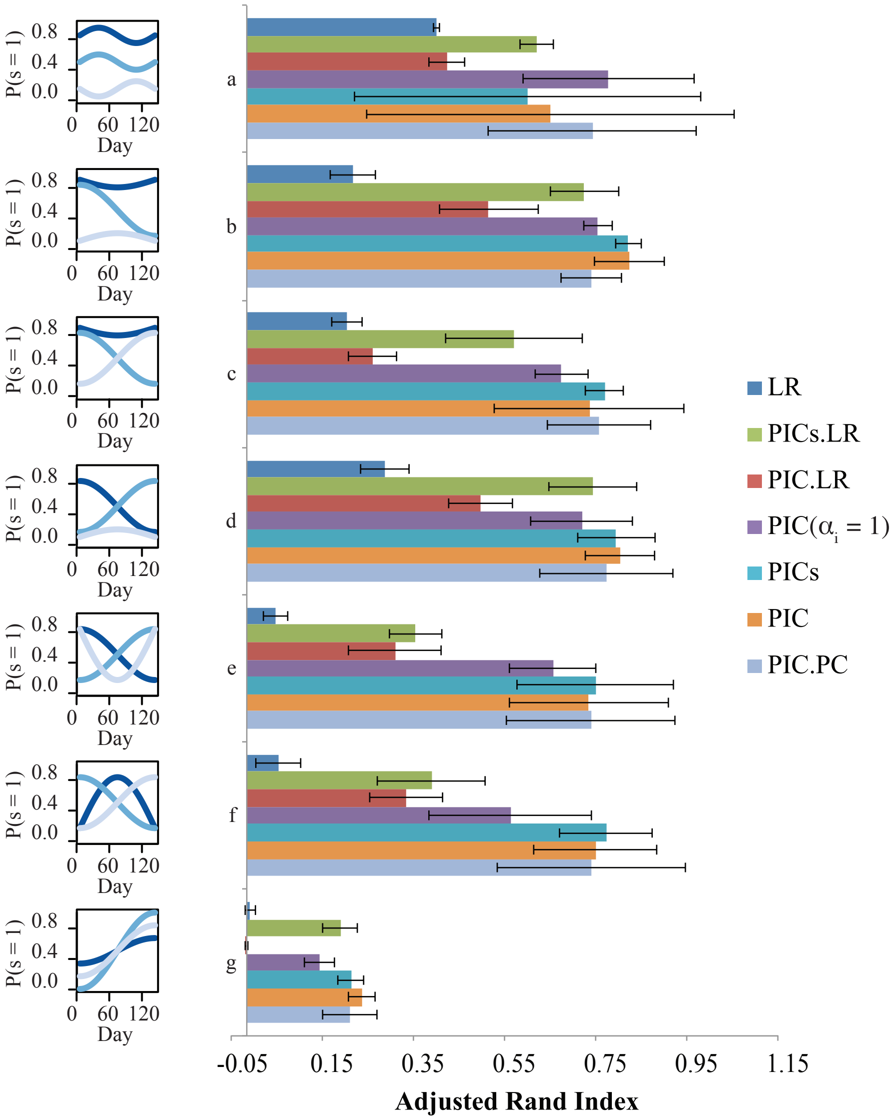


Figure 1. Simulation results for Setting 1, where we consider 8 different sets of symptom frequency profiles while fixing Δ = 0.3, and using r_i = 0.4 for all users. Symptom profiles (a-g, left) are defined for sets of 3 clusters, where each cluster is color-coded from highest (dark) to lowest (light blue) overall mean symptom rate. Average adjusted rand indices and their standard deviations across 50 simulations with 100 iterations of imputation each are shown for all strategies.

The strategies involving low rank matrix completion display more variability across symptom profiles, particularly the LR strategy which shows a clear decrease in performance as simulation scenarios become more difficult. The accuracy of all methods tend to decrease with the overall missing rate of the data (Fig. 2A). LR is particularly worse in cases where the overall non-response rate (r_i) or the severity level of non-random missing (Δ) is high (Fig. 2B). We also observe disadvantages of PIC($\alpha_i = 1$) compared to PIC under these same circumstances, due to the lack of treatment of non-ignorable missing (Fig. 2A and Fig. 2B). Most strategies show comparable performance across different α_i scenarios, with the exception of LR, which shows enhanced performance when α_i is set to a2 (Fig. 2C). In the end, Fig. 2D suggests that PIC achieves better power to detect association between covariates and predicted clusters than other clustering strategies when the strength of association is simulated to be more moderate.

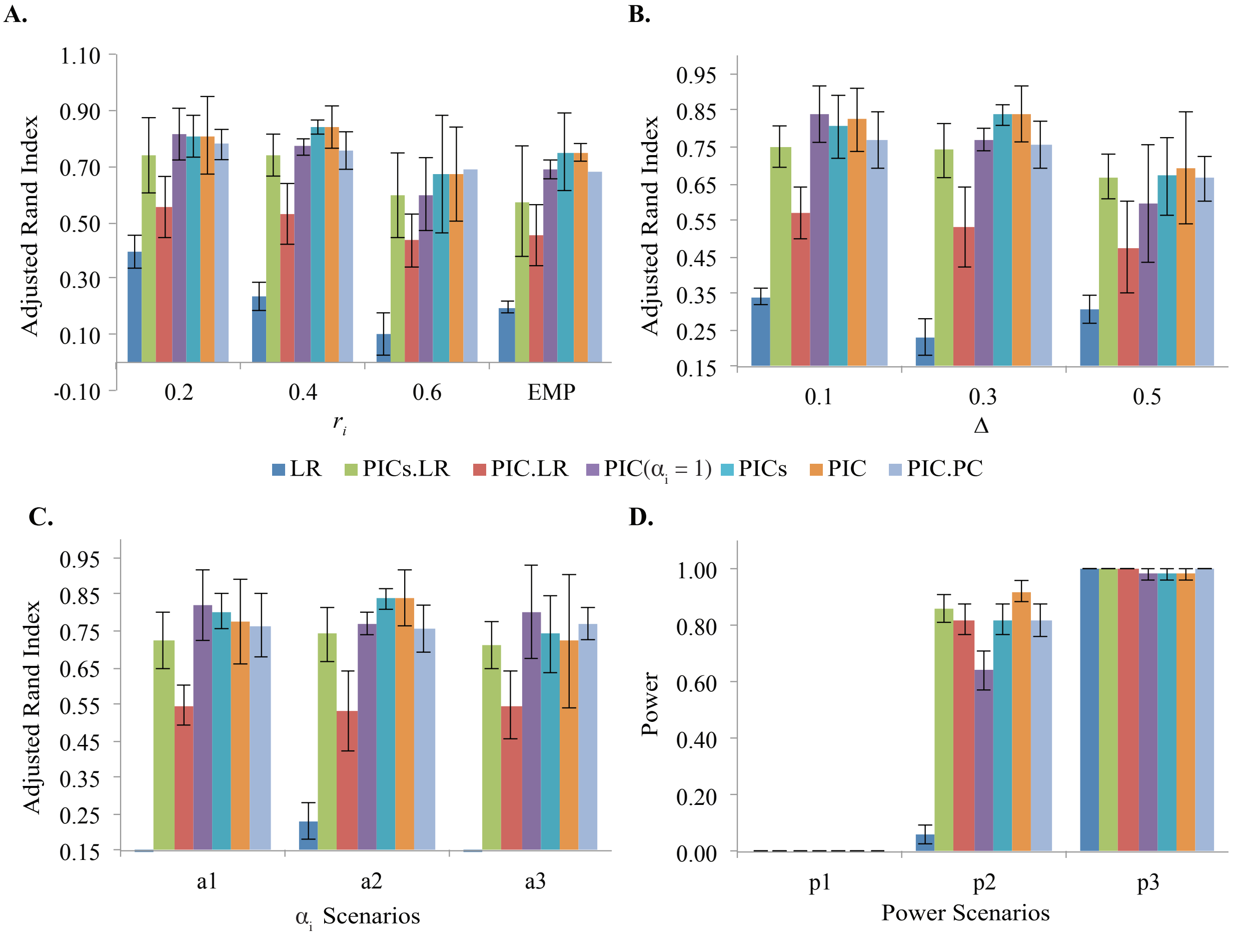


Figure 2. Simulation results based on 50 data simulations with 100 multiple imputations each. A. Results for setting 2, where we consider several values of r_i, including: 0.2, 0.4, and 0.6, where r_i is a constant for all users; and EMP, where r_i is sampled from the empirical distribution of missing rates calculated from AMHS data. B. Results for setting 3, considering several values for Δ. C. Results for setting 4, where we consider different scenarios for assigning values to the random variable α_i, where the maximum fold-difference in $P(s_{it} = 1|d_{it} = 1)/P(s_{it} = 1|d_{it} = 0)$ varies from 1.5 (a1) to 2 (a2) to 2.5 (a3). D. Power analysis based on 3 scenarios of simulated covariate data varying from null (p1) to strongest association (p3) with true cluster assignments.

4. Analysis of the AMHS data using PIC

Clustering analysis was performed for several data types, including daily symptoms and daily self-reports of asthma triggers on air quality, heat, and pollen. Study participants were first clustered into subtypes using daily symptom data collected by the AMHS. To further characterize these subtypes, we tested for associations between predicted cluster assignments and clinical variables (age of diagnosis, GINA control level, smoking status, and weight), demographic variables (gender, income, and ethnicity), as well as self-reported trigger data collected by our app (pollen, heat, and air quality). Tests of association were performed using Fisher's exact tests, where we filtered out categories with fewer than 10 individuals where applicable. Supplemental Table 1 summarizes these results (http://icahndigitalhealth.org/wp-content/uploads/2016/08/Clustering-Supplemental-Data.pdf).

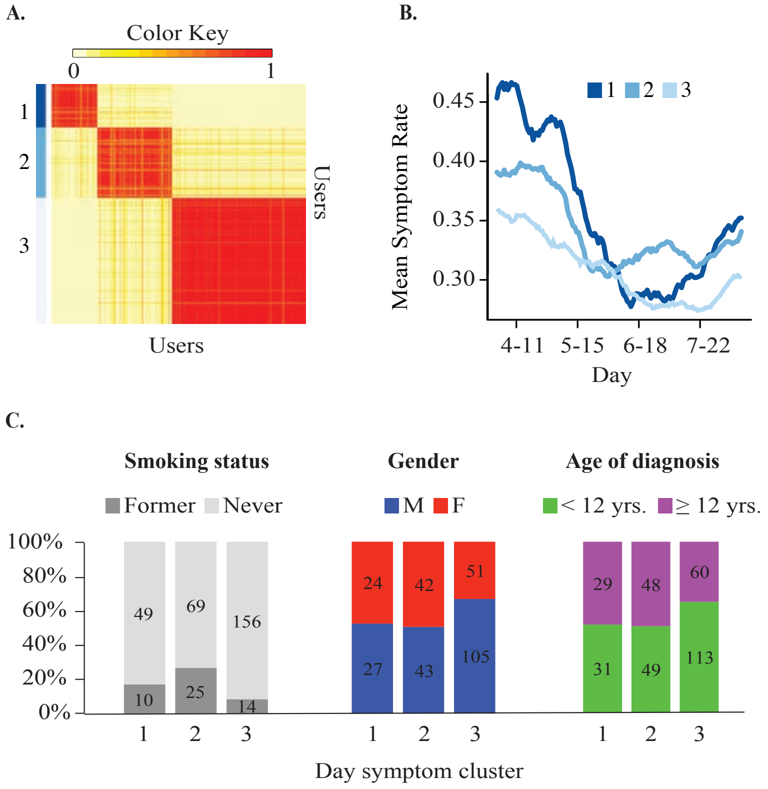


Figure 3. A. Heatmap is based on the adjacency matrix derived from consensus clustering of daily asthma symptoms for 334 users over 136 days using strategy PIC based on 100 iterations of imputation. Three distinct clusters ($n_1 = 60$, $n_2 = 98$, and $n_3 = 176$) are identified by color and enumeration (1-dark, 2-medium, and 3-light blue) where pairs of users most frequently found in the same cluster are found in the red regions along the diagonal. B. Mean curves for the clusters are based on the average of smoothed-imputed data on asthma symptoms. Each curve shows the mean symptom rate for users belonging to each cluster. Clusters are color-coded from dark to light blue by the overall mean symptom rate for each cluster. C. Day symptom clusters are significantly associated with smoking status (p = 0.0005), gender (p = 0.02), and age of diagnosis (p = 0.03) based on Fisher's exact test with simulated p-values based on 2,000 replicates. Barplots show the percentage distribution for each category within each day symptom cluster.

To conduct clustering analysis, we considered daily survey data collected by the app over the 6-month period from March 9, 2015 through August 9, 2015. We restricted our analysis to nonsmokers, defined as either having never smoked or having smoked less than 10 packs per year, without congestive heart failure or lung diseases other than asthma. We further required that each user have at least 50 survey responses over the entire 6-month period. These filterings led to 334 users in total. To ensure adequate overlap among enrollment periods across these users for comparing among methods in our simulation studies, we restricted our analysis to the 136-day period from April 2, 2015 (early spring) to August 8, 2015 (late summer), which corresponds to 136 days in total.

Based on daily survey data from 334 users over a period of 136 days, the average number of surveys provided per user was 70 (SD = 25), with an average per user enrollment period of 109 days (SD = 25). The average within enrollment missing rate was .4 (SD = 0.2). Clustering on the daily asthma symptom data was performed using the PIC strategy. After running the PIC method separately using different cluster numbers ranging from 2 to 5, we determined that users were well grouped into 3 clusters based on visual comparison of heatmaps derived from the adjacency matrices produced during the consensus clustering step of each run (Fig. 3A). Mean curves based on the average symptom rate for the users belonging to each of these clusters is shown in Figure 3B based on the average of the smoothed imputed data across 100 iterations of imputation, where curves are color-coded from dark to light blue to identify clusters with high, middle, and low symptom rates based on averaging across days.

We first sought to characterize our derived day symptom subtypes by comparing them with clinical and demographic variables. We found a significant association between asthma symptoms and smoking status (Fisher's exact test: p = 5e-4; n = 333), gender (Fisher's exact test: p = 0.02; n = 292), and age of diagnosis (Fisher's exact test: p = 0.03; n = 330). To study the relationship between asthma subtypes and environmental triggers, we used a similar approach to cluster self-reported data on daily asthma triggers collected by the AMHS. In the daily survey, participants were asked to self-report on symptom triggers on a given day. Specifically, users were able to choose from a list of 22 known asthma triggers, including allergens such as pollen, pet dander, and weather conditions. We chose to focus our analysis on air quality, heat, and pollen trigger data based on results from previous validation efforts comparing trigger data with more objective measures (PM2.5, max daily temperature, and pollen counts) using publicly available datasets[1].

Triggers were coded as 0/1 depending on whether a user cited a given trigger on a given day. Although we know that missing data in symptom reports were not random, we have little basis for attributing non-reported symptoms to one trigger over another with greater probability. Therefore, in conducting missing data imputation for trigger data, we used PIC($\alpha_i = 1$). Heatmaps resulting from the application of this method are shown in Supplemental Figure 1A-C (http://icahndigitalhealth.org/wp-content/uploads/2016/08/Clustering-Supplemental-Data.pdf).

Based on these groupings, self-reported asthma triggers were associated with the day symptom cluster groupings. Specifically, with Fisher's exact test, we found highly significant associations between day symptom clusters and clusters derived from self-reported data on pollen (p = 5e-4; N = 333), heat (p = 5e-4; N = 333), and air quality (p = 0.02; N = 333) triggers. As expected, we found a significant association between heat and US climate regions[13] broken down by northern and southern regions (Supplemental Table 2), with users belonging to cluster H1, who reported peak heat trigger complaints in late July, more frequently located in the northern US climate regions (72%) (p = 0.01; N = 288). We found that asthma trigger clusters differentiated by asthma subtype such that users who complain most frequently of pollen and heat are most frequently found in day symptom cluster 1, corresponding to the group with the highest average day symptom levels (Fig. 4A-B). By contrast, individuals frequently citing air quality as their asthma trigger are more frequently found in cluster 3, corresponding to the lowest overall day symptom rate.

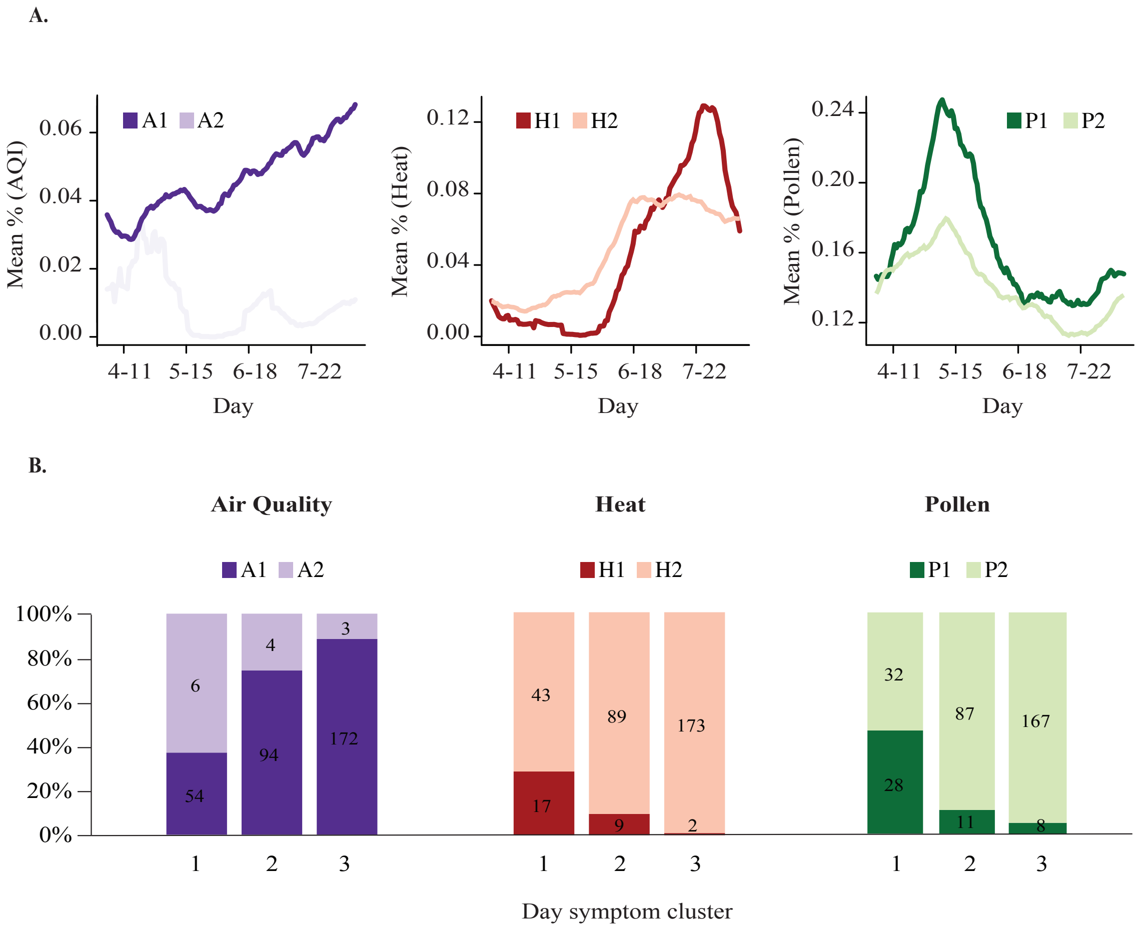


Figure 4. A. Curves depict the mean percentage of users reporting air quality, heat, and pollen for each cluster derived from the application of PIC($\alpha_i = 1$) using 100 multiple imputations. Clusters are color-coded from dark (high) to light (low) according to the overall mean percentage for each cluster averaged across days. B. Day symptom clusters are significantly associated with trigger clusters for air quality (p = 0.02, N = 333), heat (p = 5e-4, N = 333), and pollen (p = 5e-4, N = 333), based on Fisher's exact test with simulated p-values based on 2,000 replicates.

5. Discussion

Here we have considered the problem of clustering time series data collected from mobile health apps in which there is a high proportion of missing data for which the missing data mechanism is at least partially known. For such cases, regular clustering methods cannot be applied directly. To bridge this gap, in this paper, we developed an integrated PIC strategy to both impute the missing data using a probabilistic model and then clustered samples to identify subgroups with distinct patterns. The advantage of our PIC approach over other strategies based on low-rank matrix completion is demonstrated through extensive simulation studies.

When applying PIC on the AMHS data, we identified a unique subgroup of patients who have relatively high symptom rates and are more sensitive to distinct environmental factors with seasonal changes, such as heat and pollen. Furthermore, we noted relatively lower reported symptom rates associated with air quality, which may be attributed to the multi-factorial, reduced variability, and less well defined nature of this asthma trigger. With further validation, the ability to identify unique disease patterns in data sets with non-random missing data could be extremely useful in the conduct of environmental epidemiologic research as it could be used to track and identify novel environmental risk factors linked to worsening asthma. Moreover, it could enable us to identify at risk populations in large data sets and design targeted interventions to apply to reduce risk and improve outcomes. The ability to monitor asthma symptoms longitudinally by mobile technology, and identify specific subgroups of patients who have destabilization of asthma control based on specific triggers creates the opportunity to intervene early therapeutically. For example, if high heat or high pollen conditions are identified using personalized reports available by mobile technology, personalized alerts regarding presence of triggers would allow patients to seek medical advice and potentially adjust therapy in order to avoid the need for urgent care. R code implementing PIC (probability based imputation and consensus clustering) can be found here: http://icahndigitalhealth.org/wp-content/uploads/2016/10/PIC.R.

6. References

1. Chan, Y.-F.Y., et al., *The Asthma Mobile Health Study, a Large Scale Clinical Study Using ResearchKit.* Nature Biotechnology, submitted., 2016.
2. *Asthma-Data, Statistics, and Surveillance: Center for Disease Control and Prevention* 2015.
3. *GINA guidelines: Global Initiative for Asthma*. 2016.
4. Gauthier, M., A. Ray, and S.E. Wenzel, *Evolving Concepts of Asthma.* American Journal of Respiratory and Critical Care Medicine, 2015. **192**(6): p. 660-668.
5. Kaminsky, D.A., *Systems biology approach for subtyping asthma; where do we stand now?* Current opinion in pulmonary medicine, 2014. **20**(1): p. 17-22.
6. Chung, K.F., *Defining phenotypes in asthma: a step towards personalized medicine.* Drugs, 2014. **74**(7): p. 719-728.
7. Rubin, D., *Inference and missing data.* Biometrika 63 (3), 581-592, 1976.
8. Rubin Donald, B., *Multiple imputation for nonresponse in surveys*. 1987, New York: Wiley.
9. EJ Candès, B.R., *Exact matrix completion via convex optimization.* Foundations of Computational Mathematics 9 (6), 717-772.
10. Filkov, V. and S. Skiena, *Integrating microarray data by consensus clustering.* International Journal on Artificial Intelligence Tools, 2004. **13**(04): p. 863-880.
11. Hastie, T. and R. Mazumder, *softImpute: Matrix Completion via Iterative Soft-Thresholded SVD.* R package version, 2015. **1**.
12. Hubert, L. and P. Arabie, *Comparing partitions.* Journal of classification, 1985. **2**(1): p. 193-218.
13. Karl, T. and W.J. Koss, *Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983*. 1984: National Climatic Data Center.

A NEW RELEVANCE ESTIMATOR FOR THE COMPILATION AND VISUALIZATION OF DISEASE PATTERNS AND POTENTIAL DRUG TARGETS

MODEST VON KORFF

Research Information Management, Actelion Pharmaceuticals Ltd., Gewerbestrasse 16
Allschwil, 4123, Switzerland
Email: modest.korff@actelion.com

TOBIAS FINK

Research Information Management, Actelion Pharmaceuticals Ltd., Gewerbestrasse 16
Allschwil, 4123, Switzerland
Email: tobias.fink@actelion.com

THOMAS SANDER

Research Information Management, Actelion Pharmaceuticals Ltd., Gewerbestrasse 16
Allschwil, 4123, Switzerland
Email: thomas.sander@actelion.com

A new computational method is presented to extract disease patterns from heterogeneous and text-based data. For this study, 22 million PubMed records were mined for co-occurrences of gene name synonyms and disease MeSH terms. The resulting publication counts were transferred into a matrix $\mathbf{M}_{\text{data}}$. In this matrix, a disease was represented by a row and a gene by a column. Each field in the matrix represented the publication count for a co-occurring disease–gene pair. A second matrix with identical dimensions $\mathbf{M}_{\text{relevance}}$ was derived from $\mathbf{M}_{\text{data}}$. To create $\mathbf{M}_{\text{relevance}}$ the values from $\mathbf{M}_{\text{data}}$ were normalized. The normalized values were multiplied by the column-wise calculated Gini coefficient. This multiplication resulted in a relevance estimator for every gene in relation to a disease. From $\mathbf{M}_{\text{relevance}}$ the similarities between all row vectors were calculated. The resulting similarity matrix $\mathbf{S}_{\text{relevance}}$ related 5,000 diseases by the relevance estimators calculated for 15,000 genes. Three diseases were analyzed in detail for the validation of the disease patterns and the relevant genes. Cytoscape was used to visualize and to analyze $\mathbf{M}_{\text{relevance}}$ and $\mathbf{S}_{\text{relevance}}$ together with the genes and diseases. Summarizing the results, it can be stated that the relevance estimator introduced here was able to detect valid disease patterns and to identify genes that encoded key proteins and potential targets for drug discovery projects.

I. Introduction

Many diseases coexist in a biological context with other diseases [1]. Patients often suffer from more than one disease. Furthermore, it is well known that some diseases cause secondary diseases. A well-established example for a disease with many co-occurring and second-order diseases is diabetes mellitus [2]. In addition, the context of a disease is of crucial importance in drug discovery. The ultimate goal of any new project in drug discovery is to treat or cure the disease in question. Realistically, however, for truly innovative drug discovery projects, in which the selected targets have been only recently identified, only sparse knowledge is available about the relationship between the chosen target and the potential disease. Knowledge about the context of the disease in which the target is involved may help to decide which constellation of diseases to take into account. Later in the drug discovery process, coexisting diseases should be considered for toxicity and DMPK studies. Drugs that are used for the treatment of pre-existing conditions may influence the metabolism of the patient and may result in an interaction with the drug being tested.

Whereas earlier research approaches for studying coexisting diseases were mainly based on phenotypic observations, recent technical advances have paved the way for using genomic and proteomic data. In this context, the “Online Mendelian Inheritance in Man”(OMIM) database was first published as a book and later as an electronic database [3]. This database related genotypes to phenotypes and inspired several research groups to develop computational tools to derive disease–disease associations. One of the first disease–disease association tools was reported by Goh et al. [4]. They developed a human disease network, connected by genes that were associated with two diseases. An enrichment of disease candidate genes via text mining of OMIM descriptions was implemented by van Driel et al. [5]. MeSH (Medical Subject Headings) annotations of MEDLINE articles were analyzed by Liu et al. to extract genetic and environmental factors associated with certain diseases [6]. An overview of the current research in disease associations was recently published by Sun et al. [7]. Hidalgo et al. derived the “Phenotypic Disease Network”(PDN) from 32 million patient records and received about six million co-morbidity relations [8]. This disease association network contains the co-morbidity data for more than 10,000 ICD9-encoded diseases and is one of the largest known so far. Taken together, a review of the existing literature suggests that multiple approaches exist to derive disease associations. However, MEDLINE represents the largest source of data, but it has not been used exhaustively for deriving disease–disease associations so far.

The approach presented here—the Disease–Disease Relevance Miner DDRelevanceMiner—annotated all records in MEDLINE with gene names and MeSH terms. Disease–disease associations were derived by comparing gene name based word vectors. These word vectors are histograms which are extracted from the records in PubMed by mining for co-occurrences of gene names and disease terms. This technique is known as second order co-occurrence and has been used by Schütze for word sense discrimination [9]. He exploited the fact that similar words are often accompanied by groups of identical words. Second-order co-occurrence has the advantage, as it allows calculating the similarity between two words that do not co-occur frequently, but that co-occur with the same neighboring words. Our method differs from Schütze's approach in two ways. The DDRelevanceMiner creates a word vector for a disease term from the complete text corpus and not from a single record. Word vectors, which represent a single text record, are often normalized and then weighted by the inverse document frequency. Weighing by the inverse document frequency is not feasible for DDRelevanceMiner, because each word vector contains counts from all text records. To overcome this problem we introduced the relevance estimator.

In the next section, the calculation of the relevance estimator is explained in detail. Additionally, the data are described which were used to feed the algorithm and the diseases which were chosen for a thorough test of the results. The presentation and a discussion of the results follow at the end of the manuscript.

II. METHODS

A detailed description of the algorithms used by the precursor of DDRelevanceMiner—DDMiner—has been recently published [10]. Here, we describe the new algorithm which significantly improved the results of DDMiner. The new algorithm calculates a relevance estimator for a gene in dependency to a disease. How can one assess the merits of the relevance estimators? We assumed that ranking genes by their relevance estimators should help identifying potential drug targets. We also assumed that calculating similarities between diseases based on the relevance estimators should group these diseases in a meaningful way. Finally, if the relevance estimators are applied to a well-studied disease, it should be possible to prove the importance of the top-ranked genes and the disease patterns by literature. The relevance estimator is loosely related to the scoring scheme that was recently published by Mørk et al. [11].

A. Description of DDRelevanceMiner

DDRelevanceMiner used for analysis all available gene names from a table provided by the HUGO Gene Nomenclature Committee (HGNC) [12]. Gene name synonyms were taken from the HUGO table and other public available sources [13;14]. Every synonym was checked for ambiguity. Every synonym that passed the check was used to form a query for PubMed. A successful query retrieved a number of PubMed records. Index, title, and abstract of the record were searched for disease MeSH terms. All found disease MeSH terms were labeled with the approved gene symbol that was linked to the PubMed record. A detailed description of the search algorithm for gene and disease terms has been previously described in [10].

Querying PubMed with all gene name synonyms and parsing all retrieved records with all disease MeSH terms resulted in the central data matrix $\mathbf{M}_{\text{data}}$. A row in the matrix stood for a disease MeSH term and a column – for an approved gene symbol. Each field in the matrix, indicated by a row and a column, contained an integer number indicating how often a disease MeSH term occurred together with a gene. A row *m* from $\mathbf{M}_{\text{data}}$ shows which genes were studied together with the disease *m*. Vice versa, a column *n* shows which diseases were reported together with gene *n*.

However, the pure count of disease–gene co-occurrences is only of limited benefit. Genes with a high frequency of occurrence in the medical literature are often studied in relation to many diseases. But pharmaceutical research is mostly interested in genes that are specific for the disease of interest. Most interesting are genes that are specifically mentioned together with a disease of interest and not together with other genes. A gene with a high number of occurrences with one disease and no mentioning together with other diseases could be assumed to have a high relevance for the disease. Column *n* was extracted from $\mathbf{M}_{\text{data}}$ to calculate the relevance estimator $r_{\text{m,n}}$ for a disease with index *m* and a gene with index *n*. From column *n*, only fields with a publication count >0 were considered. For all fields in the column, their rank fraction $f_{\text{m,n}}$ was calculated. The rank ρ of a disease *m* for gene *n* is the position of the disease after sorting column *n* according to the number of publications. Diseases with identical publication counts were assigned the same rank.

Consequently, the number of ranks can be smaller than the number of diseases that were mentioned together with gene *n*. The rank fraction equaled one minus the rank divided by the total number of ranks $f_{m,n}= 1-\rho/\theta$, with θ for the total number of ranks. A fraction of publications $p_{m,n}$ was calculated by dividing the number of publications for disease *m* found in column *n* by the sum of all publications for gene *n*. The fraction of publications for disease *m* was weighted with the relative rank by $w_{m,n}= p_{m,n} f_{m,n}$. Finally, the relevance estimator $r_{m,n}$ was calculated by multiplying $w_{m,n}$ by the Gini coefficient g_n. The Gini coefficient describes the statistical dispersion for a group of values [15]. A Gini coefficient close to one indicates that all values except one in the group are zero. A Gini coefficient of zero indicates that all values in the group are equal. A relevance estimator $r_{m,n}$ of one is obtained if all three factors in the equation $r_{m,n}= p_{m,n} f_{m,n}\, g_n$ are equal to one. This means that all publications for gene *n* refer only to disease *m*. The calculation of the relevance estimator was done for all fields in the matrix $\mathbf{M}_{\text{data}}$. As result, a new matrix $\mathbf{M}_{\text{relevance}}$ was obtained, with the same dimensions as the input matrix. This matrix contained the relevance estimators; they covered a range between zero and one. This normalization enabled the meaningful comparison of matrix rows. To reduce the risk of rounding errors and to cut off the influence of very small values, the relevance estimators were multiplied by a factor of 1,000 and converted into integer numbers. Each two $\mathbf{M}_{\text{relevance}}$ matrix rows were compared by calculating the generalized Jaccard similarity coefficient. Comparison of two matrix rows gave a similarity value between two diseases. The resulting similarity matrix $\mathbf{S}_{\text{relevance}}$ contained the similarity between all diseases.

B. *Data*

1) *Genes and disease MeSH terms*

A total of 39,410 approved gene and protein symbols were retrieved from the HUGO table. At least one disease MeSH term was found for 15,203 approved gene symbols. This number defined the number of columns in $\mathbf{M}_{\text{data}}$ and $\mathbf{M}_{\text{relevance}}$. A number of 5,256 unique MeSH descriptors defined the number of rows in $\mathbf{M}_{\text{data}}$ and $\mathbf{M}_{\text{relevance}}$.

2) *Example diseases to assess the quality of the relevance estimators*

Three example diseases, type 2 diabetes mellitus (T2DM), melanoma, and vitiligo were chosen for a detailed analysis of the disease–gene and the disease–disease associations. For each of the three diseases, five genes with the highest relevance estimators—the top genes—and the equal number of genes with the highest publication counts were analyzed. Additionally, for each of the three example diseases, ten most similar diseases were evaluated. Based on the working hypotheses described at the beginning of the Methods section, the following success criteria were defined. The usage of the relevance estimator can be regarded as a success if the top five genes include disease relevant genes. These genes should not be related to numerous other diseases. A relation of an example disease to a similar disease was regarded as valid if there was evidence found in literature.

T2DM is a subtype of diabetes and a complex metabolic disease. It is a complex disease because it involves environmental factors and multiple genes [16;17]. Despite the fact that many anti-diabetic drugs are on the market, the need for new anti-diabetic drugs is still high [18]. Obesity is a dominant risk factor for T2DM, whereas hypertension is one of the main co-occurring diseases. Therefore, we expected to see at least these two disease MeSH terms among the results of the similarity analysis. A number of genetic-driven studies were done for T2DM. Specific genes were found that increase the risk for this disease. Will the relevance estimator be able to identify some of these genes?

Melanoma is a malignant neoplasm (cancer) of the skin and the leading cause of death due to skin disease [19]. It is considered a highly immunogenic tumor [20]. Consequently, we expected to see association between melanoma and the genes that are tumor related but also with the genes that have relevance for the immune response.

Vitiligo is a disease where parts of the skin lose their pigment. Vitiligo is a frequently occurring disease, seen in 0.2%–2% of the population. However, its cause is still unknown [21]. A reason to choose vitiligo as an example disease was the relatively small number of related publications, compared to T2DM and melanoma. Another reason was the existing link between melanoma and vitiligo [22]. Will the disease similarity analysis based on the relevance estimator be able to find the link between these two diseases?

III. RESULTS

After querying PubMed with the synonyms from 39,410 approved gene and protein symbols, 2.7 million unique PubMed records were retrieved. Each of these records contained at least one disease MeSH term together with an unambiguous gene name synonym. The number of rows in $\mathbf{M}_{\text{data}}$ and $\mathbf{M}_{\text{relevance}}$ were defined by 5,256 unique disease MeSH terms found in the above publications, and the number of columns – by the 15,203 approved symbols. The similarity matrix $\mathbf{S}_{\text{relevance}}$ was calculated as described above in Methods. All results described in the following paragraphs were taken from $\mathbf{M}_{\text{data}}$, $\mathbf{M}_{\text{relevance}}$ and $\mathbf{S}_{\text{relevance}}$. The results for all genes and diseases are freely accessible at http://gene2disease.org.

A. *Results for the three example diseases*

Table 1 summarizes the disease–gene associations for the three example diseases. The table contains the disease name, the number of publications for the disease, the total number of genes found in the publications, and the top ten approved gene symbols. In the fourth column, the approved gene symbols are sorted by the relevance estimator from $\mathbf{M}_{\text{relevance}}$. In the fifth column, gene symbols are sorted by their publication counts from $\mathbf{M}_{\text{data}}$.

For every disease, the ten most similar diseases were extracted from $\mathbf{S}_{\text{relevance}}$. Cytoscape was used to visualize the results [23]. A Cytoscape network was created for every example disease (Figs. 1–3). A Cytoscape sketch for a disease contains the results from the corresponding row of Table 1 and from the corresponding table with the similar diseases. Every gene in a sketch is connected to the example disease. A disease is never directly connected to another disease and a gene is never connected directly to another gene. The example disease is marked by white text on the label. Other diseases are depicted as rectangles with black text on the label. The background color of the disease label corresponds to the similarity with the example disease. Similar diseases have a more intense background color than less similar ones. Genes are represented by ellipsoids or diamond shapes. The width of each shape corresponds to the number of connected diseases. A diamond shape symbolizes a gene that is in the top-five list of the genes sorted by relevance in Table 1. An ellipsoid indicates a gene that has similar relevance for two or more diseases, including an example disease, and is among the top five genes for one of the similar diseases.

1) *Type 2 diabetes mellitus*

DDRelevanceMiner found about 31,000 publication records relevant to T2DM which collectively mentioned synonyms for 2,273 genes (Table 1). All five genes with the highest relevance estimators have demonstrated relevance to T2DM. Namely, TCF7L2 is a diabetes susceptibility gene of substantial importance [24], and a potential target for anti-diabetic drugs [25]. GCK (glucokinase) plays an important role in the carbohydrate metabolism and is an attractive target for new drugs [18], whereas GCK mutations are known to cause diabetes. CAPN10, SLC5A2, and SLC30A8 have high relevance for T2DM and are potential drug targets.

Ranking of the same set of genes by publication count differed completely from that by relevance estimators. Four out of the five genes with the highest publication counts, GCG (glucagon), INS (insulin), HBA1 (hemoglobin) and DIANPH, had low relevance estimators. This means that they were mentioned together with many other diseases, i.e., are not specific for T2DM. Low relevance of INS to T2DM was expected, because T2DM is a non–insulin-dependent disease [26]. In contrast, the fifth gene, DPP4 (dipeptidyl peptidase-4), had both a high publication count and a relatively high relevance estimator and is, indeed, a proven drug target for treating T2DM [27].

Table 1. Three selected example diseases and top relevant genes.

Disease	Publications	Gene count	Approved gene symbol (relevance estimator, publications) Top five genes sorted by	
			relevance	publication count
Type 2 diabetes mellitus	31,024	2,273	TCF7L2 (0.140, 386) GCK (0.138, 767) CAPN10 (0.132, 106) SLC5A2 (0.123, 210) SLC30A8 (0.120, 99)	GCG (0.053, 3761) INS (0.036, 3465) HBA1 (0.090, 2374) DIANPH (0.040, 2352) DPP4 (0.113, 1937)
Melanoma	27,000	3,271	MAGEA11(0.236, 14) TYR (0.209, 2357) PMEL (0.206, 23) MIA (0.202, 138) DCT (0.196, 253)	TYR (0.209, 2357) IL2 (0.018, 2191) IFNA1 (0.010, 2007) IFNG (0.010, 1867) IFNA2 (0.010, 1261)
Vitiligo	1,608	380	PCBD1 (0.033, 3) DCT (0.017, 28) PMEL (0.016, 3) LRR1 (0.015, 4) TYR (0.013, 170)	TYR (0.013, 170) CAT (0.0004, 70) IFNG (0, 52) IL2 (0, 49) IFNA1 (0, 46)

Table 2 lists the disease MeSH terms most similar to T2DM based on $\mathbf{S}_{\text{relevance}}$. All ten MeSH terms are known as major co-occurring diseases with diabetes, symptoms of diabetes, or, in case of obesity and body weight, known risk factors for the disease [28]. The size of the common gene set linking each disease with T2DM ranged from 72 to 517 genes. The 'top five genes' column of Table 2 shows genes whose relevance estimators calculated for the disease/MeSH term were most similar to those calculated for T2DM. For this reason, the genes and their order is different from Table 1 showing genes most relevant to T2DM only.

In Table 2, the disease MeSH terms most similar to T2DM are listed. The similarity values were taken from $\mathbf{S}_{\text{relevance}}$. All ten MeSH terms are major co-occurring diseases with diabetes, symptoms of diabetes, or, in case of obesity and body weight, known risk factors for the disease [28]. The size of the common gene set linking each disease with T2DM ranges from 72 to 517 genes. The top five genes are the genes with the relevance estimators that are most similar to those calculated for T2DM. For this reason, there is a difference in the sorting order by the relevance estimators compared with Table 1, where the top genes were those most specific for one disease. Genes that connect more than one disease with T2DM are, e.g., SLC2A4, GCK, PLTP, and APOB. These genes may be regarded as having a key role in the disease pattern. The importance of these key genes is visualized in Fig. 1 where the width of the gene nodes corresponds to the number of connections.

Table 2. Diseases most similar to type 2 diabetes mellitus.

Disease/MeSH term	Similarity	Size of the common gene set	Top five genes
Insulin Resistance	0.81	484	SLC2A4, IRS1, INSR, CAPN10, IRS2
Hyperglycemia	0.79	253	SLC5A2, GCK, PDX1, GCGR, HBA1
Obesity	0.72	517	ADIPOQ, GIP, ADRB3, FTO, GLP1R
Body Weight	0.70	286	GCK, IAPP, SLC2A4, GLP1R, UCP1
Hyperinsulinism	0.66	145	ABCC8, KCNJ11, GCK, SLC2A4, INSR
Hypoglycemia	0.61	72	SLC5A2, DPP4, GLP1R, ABCC8, HBA1
Atherosclerosis	0.58	243	PLTP, PON2, CETP, APOB, APOA1
Dyslipidemias	0.57	122	CETP, APOB, PPARA, PLTP, HMGCR
Hyperlipidemias	0.56	122	VLDLR, APOB, LPL, LIPC, PLTP
Hypertension	0.53	281	DIANPH, HBA1, ACE, ADRB3, REN

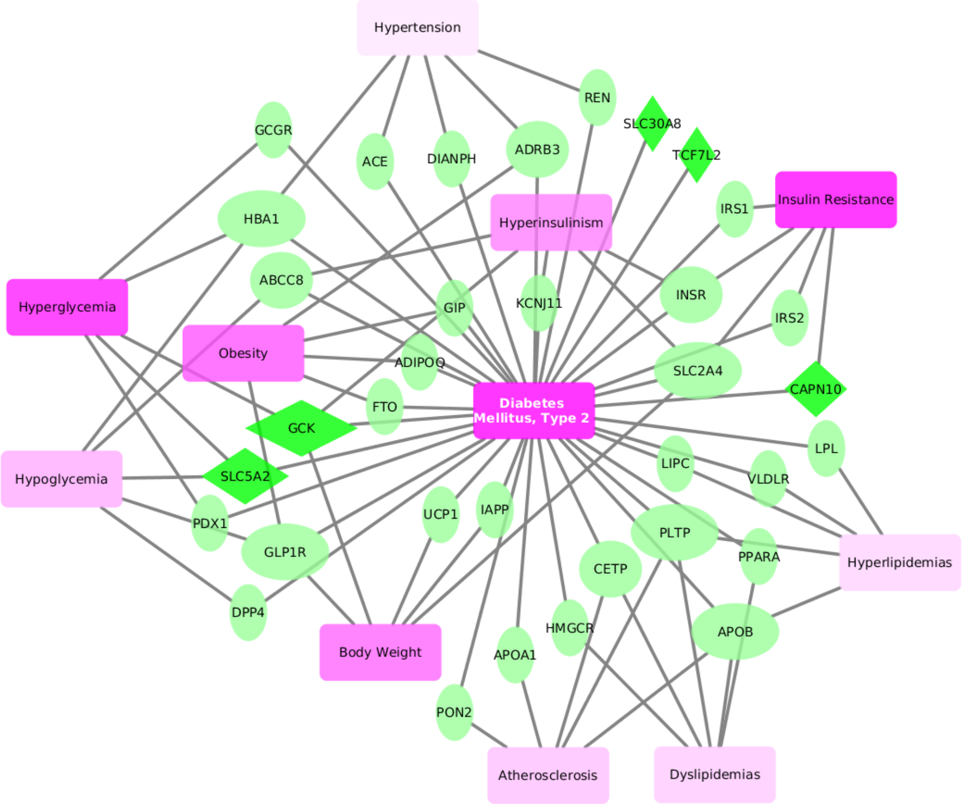


Fig. 1. Disease and gene pattern for type 2 diabetes mellitus

Analysis of gene-disease pattern for T2DM showed that SLC5A2 gene connects T2DM with two out of ten diseases with the highest similarity to T2DM (Fig. 1), which makes SLC5A2 an interesting drug target [29]. CAPN10 connects T2DM with insulin resistance, which is one of its known pre-conditions [30]. Other genes that connect T2DM with more than one disease (e.g., SLC2A4, GCK, PLTP, APOB), may be regarded as having a key role in the gene-disease pattern. The importance of these key genes is highlighted in Fig. 1 by the width of the gene nodes that is proportional to the number of disease connections.

2) *Melanoma*

Melanoma is a malignant cancer of the skin. The development of cancer includes many genes for cell growth and proliferation. Tyrosinase (TYR) is the gene with the highest publication count, and the top second rank according to the relevance estimator calculated for melanoma. Tyrosinase plays a central role in the process of skin pigmentation. Next four genes with high publication counts, interleukin 2 (IL2) and three interferons, are important for the immune response against cancer but are not specific for melanoma. The lack of specificity is the reason why these genes have low relevance estimators. MAGEA11 is the gene with the highest relevance estimator in the complete examination. Indeed, it is a melanoma antigen. Other genes with high relevance estimators, PMEL, MIA, and DCT, are highly specific for melanoma and are in the focus of ongoing research [31] [32] [33].

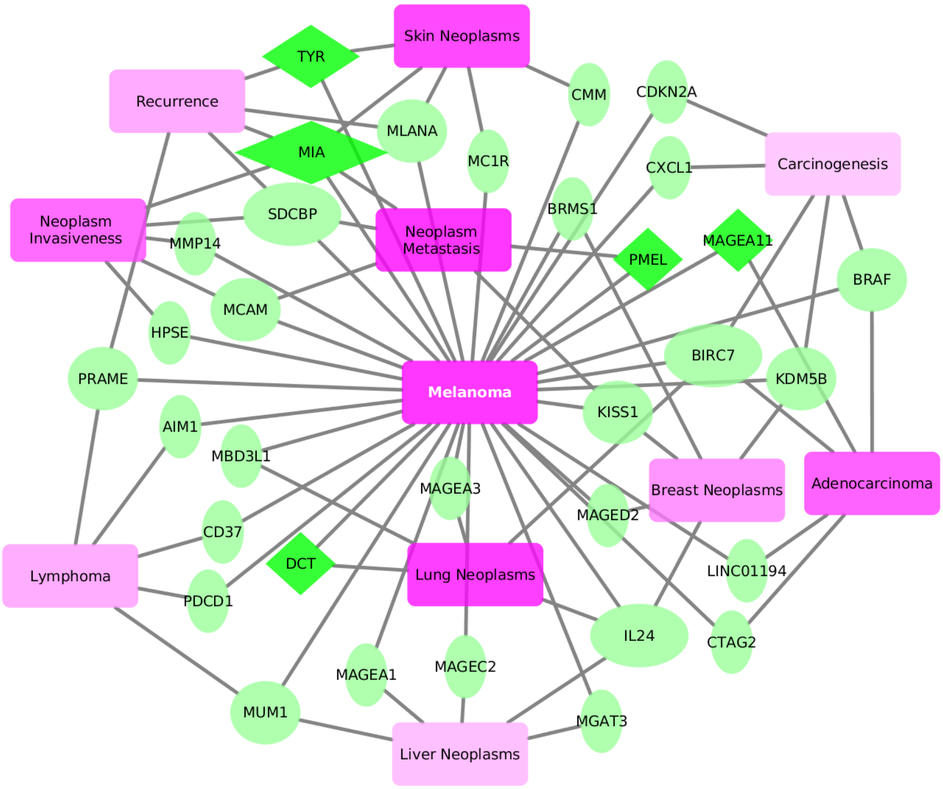


Fig. 2. Disease and gene pattern for melanoma.

The list of similar diseases in Table 3 is determined by diseases generally related to skin cancer and other cancer types. As already mentioned, the number of genes involved in cancer is higher than for other diseases. In the similarity list with the top five genes MIA the only gene that is represented four times. MIA encodes the melanoma-derived growth regulatory protein. The combination of high relevance estimator and the connection of three melanoma-related MeSH terms suggest that MIA is a strong candidate for a drug discovery project.

Table 3. Diseases most similar to melanoma.

Disease/MeSH term	Similarity	Size of the common gene set	Top five genes
Neoplasm Metastasis	0.65	1028	SDCBP, MIA, MCAM, KISS1, PMEL
Lung Neoplasms	0.65	622	MAGEA3, IL24, DCT, BIRC7, MBD3L1
Skin Neoplasms	0.65	366	MC1R, MIA, CMM, MLANA, TYR
Adenocarcinoma	0.63	728	LINC01194, BIRC7, MAGEA11, CTAG2, BRAF
Neoplasm Invasiveness	0.63	508	SDCBP, MIA, MCAM, HPSE, MMP14
Breast Neoplasms	0.60	788	BRMS1, IL24, KDM5B, MAGED2, KISS1
Recurrence	0.59	541	PRAME, MIA, MLANA, SDCBP, TYR
Lymphoma	0.59	581	CD37, PRAME, MUM1, AIM1, PDCD1
Liver Neoplasms	0.58	548	MAGEC2, MGAT3, MUM1, MAGEA1, IL24
Carcinogenesis	0.58	954	KDM5B, CDKN2A, CXCL1, BIRC7, BRAF

3) *Vitiligo*

Much less is known about vitiligo than for the other two example diseases. The absence of any gene with a high relevance estimator for vitiligo indicates a comparative lack of research.

Table 4 shows the most similar diseases to vitiligo. Coinciding with the low number of publication counts is the small size of the common gene sets. Nevertheless, some genes show multiple connections in the disease–gene network shown in Fig. 3. Tyrosinase has the most connections by linking nine diseases. Dopachrome tautomerase (DCT) connects seven diseases and is one of the most relevant genes for vitiligo. Also seven diseases are connected by the MITF gene encoding melanogenesis associated transcription factor, but this gene is not part of the top relevance genes. PMEL and TYRP1 genes connect six and five diseases, respectively. Fig. 3 shows that all four genes with the most connections (TYR, DCT, MITF, PMEL) relate vitiligo to the same three disease MeSH terms within the skin cancer complex: hypopigmentation, hyperpigmentation and skin neoplasms.

Table 4. Diseases most similar to vitiligo.

Disease/MeSH term	Similarity	Size of the common gene set	Top five genes
Hyperpigmentation	0.72	7	TYR, DCT, MITF, PMEL, TYRP1
Melanosis	0.69	5	TYR, ASIP, LGI3, MITF, MC1R
Melanoma, Experimental	0.63	8	DCT, TYR, PMEL, MITF, TYRP1
Microphthalmos	0.51	10	DCT, TYR, MITF, PMEL, ASIP
Hypopigmentation	0.49	3	TYR, DCT, MITF

Disease/MeSH term	Similarity	Size of the common gene set	Top five genes
Melanoma, Amelanotic	0.49	3	TYR, DCT, MLANA
Epilepsy, Partial, Sensory	0.48	1	LGI3
Skin Neoplasms	0.43	10	TYR, DCT, PMEL, STX17, MITF
Arthritis, Juvenile	0.42	3	PTPN22, NLRP1, PTPN2
Albinism, Oculocutaneous	0.42	5	TYR, PMEL, TYRP1, GCHFR, MC1R

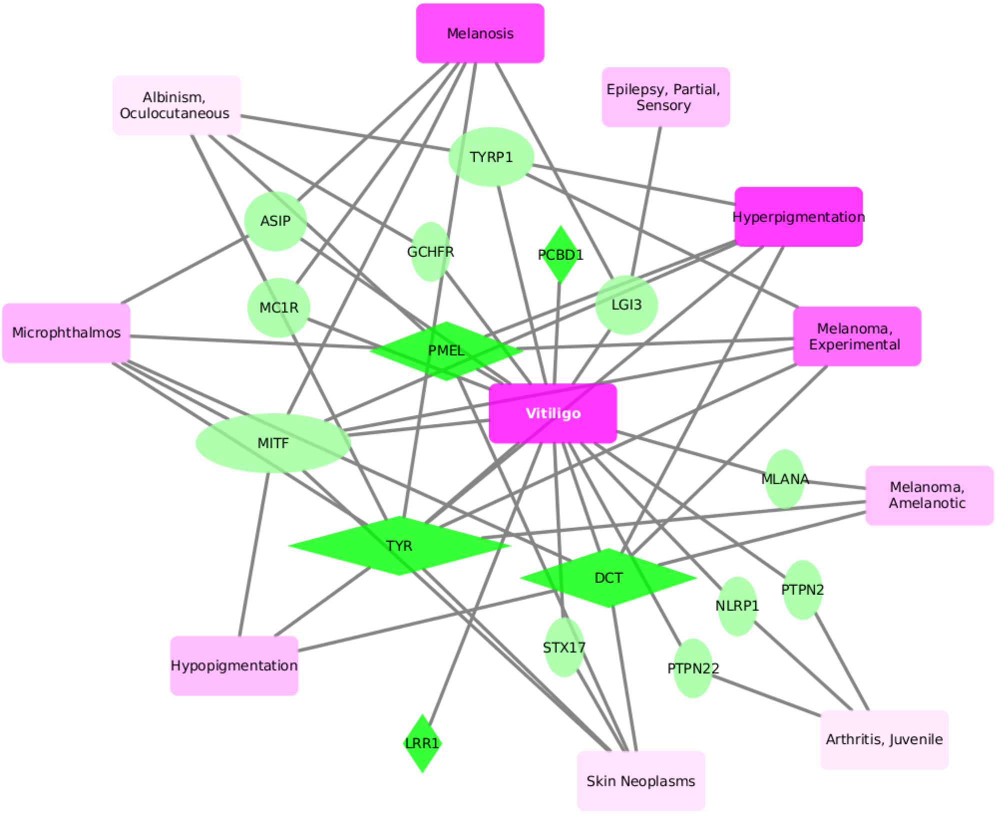


Fig. 3. Disease and gene pattern for vitiligo.

IV. Discussion

The majority of scientific publications contain information as heterogeneous data. And scientific publications are the main source for medical information. The NIH collected abstract records for millions of scientific publications in the PubMed database. It was our goal to extract and to visualize meaningful disease–gene and disease–disease patterns from this plethora of unstructured information. For this purpose, we introduced the relevance estimator described in this study. Three example diseases were chosen to examine this new figure of merit. For each disease, the relation to the most relevant genes was confirmed by literature. Furthermore, ten disease MeSH terms with disease–gene relationships most similar to one of the example diseases were identified and analyzed. Cytoscape was used to visualize the most relevant genes, the similar diseases and the genes which connect the diseases with the example disease.

The most relevant genes for T2DM and melanoma were found to be highly specific for each disease. The ability of the relevance estimator to link MeSH terms with highly disease-specific genes that may only affect small patient groups makes it interesting for personalized medicine. The gene with the highest relevance estimator, TCF7L2, has been identified as a potential anti-diabetic drug target [25]. Obesity and hypertension were among the top ten disease MeSH terms with highest similarity to T2DM. This finding was pre-defined as a success criterion for the use of the relevance estimator.

For melanoma, a different disease–gene relationship pattern was obtained than for T2DM. All top genes are connecting at least one additional disease MeSH term with melanoma. Immune system relevant genes were listed as top genes by publication counts. However, these immune regulatory genes had low relevance estimators for melanoma. An example is interleukin 2, the protein product of the IL2 gene. This protein is used as a drug in the treatment of melanoma and is known to cause adverse side effects [34]. No genes with a high relevance estimator were found for vitiligo. Here, the combination of low publication counts and low relevance estimators emphasized that vitiligo is a disease with unknown genetic causes. Regardless of the low relevance score, three of the top five genes for vitiligo connected vitiligo to other skin-related disease MeSH terms. Thus, the earlier mentioned link between vitiligo and melanoma was confirmed using relevance estimators.

The evidence provided by the relevance estimators can be summarized as follows:

1. A high relevance estimator together with a low publication count indicates potential drug targets.
2. A low relevance estimator together with a high publication count indicate non-disease-specific genes.
3. A high relevance estimator and a high publication count mark a well-studied gene that is highly specific for the related disease.
4. Genes with low relevance estimators for a certain disease and high connectivity between multiple disease MeSH terms are likely to encode key proteins in a biochemical or signaling pathway.

Concluding, the relevance estimator is a valuable tool to extract disease-gene relation patterns from very large and heterogeneous data sets. Yet, the nature and importance of these patterns can only be evaluated by a scientist.

References

1 M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, *J Chronic Dis* **40**, (1987).

2 K. G. Alberti, and P. Z. Zimmet, *Diabet Med* **15**, (1998).

3 A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick, *Nucleic Acids Res* **30**, (2002).

4 K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabasi, *Proc Natl Acad Sci USA* **104**, (2007).

5 M. A. van Driel, and H. G. Brunner, *Hum Genomics* **2**, (2006).

6 Y. I. Liu, P. H. Wise, and A. J. Butte, *BMC Bioinformatics* **10 Suppl 2**, (2009).

7 K. Sun, J. P. Goncalves, C. Larminie, and N. Przulj, *BMC Bioinformatics* **15**, (2014).

8 C. A. Hidalgo, N. Blumm, A. L. Barabasi, and N. A. Christakis, *PLoS Comput Biol* **5**, (2009).

9 H. Schütze, *Computational linguistics* **24**, (1998).

10 M. Von Korff, B. Deffarges, and T. Sander (2015). *In* "Computational Intelligence, 2015 IEEE Symposium Series on", p. 314. IEEE.

11 S. Mork, S. Pletscher-Frankild, A. Palleja Caro, J. Gorodkin, and L. J. Jensen, *Bioinformatics* **30**, (2014).

12 http://www.genenames.org

13 D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, *Nucleic Acids Res* **33**, (2005).

14 http://www.ncbi.nlm.nih.gov/gene

15 C. Gini, *Colorado College Publication, General Series* **208**, (1936).

16 L. Chen, D. J. Magliano, and P. Z. Zimmet, *Nat Rev Endocrinol* **8**, (2012).

17 R. Sladek, G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, S. Hadjadj, B. Balkau, B. Heude, G. Charpentier, T. J. Hudson, A. Montpetit, A. V. Pshezhetsky, M. Prentki, B. I. Posner, D. J. Balding, D. Meyre, C. Polychronakos, and P. Froguel, *Nature* **445**, (2007).

18 P. Gaitonde, P. Garhyan, C. Link, J. Y. Chien, M. N. Trame, and S. Schmidt, *Clin Pharmacokinet* **55**, (2016).

19 M. A. Papadakis, S. J. McPhee, and M. W. Rabow (2015). *In* "Current Medical Diagnosis & Treatment 2015", p. 101. McGraw-Hill Education.

20 T. H. Nguyen, *Clin Dermatol* **22**, (2004).

21 A. Alkhateeb, P. R. Fain, A. Thody, D. C. Bennett, and R. A. Spritz, *Pigment Cell Res* **16**, (2003).

22 K. U. Schallreuter, C. Levenig, and J. Berger, *Dermatologica* **183**, (1991).

23 P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, *Genome Res* **13**, (2003).

24 C. J. Groves, E. Zeggini, J. Minton, T. M. Frayling, M. N. Weedon, N. W. Rayner, G. A. Hitman, M. Walker, S. Wiltshire, A. T. Hattersley, and M. I. McCarthy, *Diabetes* **55**, (2006).

25 M. Ridderstrale, and L. Groop, *Mol Cell Endocrinol* **297**, (2009).

26 C. N. Hales, *Br Med Bull* **53**, (1997).

27 D. J. Drucker, and M. A. Nauck, *Lancet* **368**, (2006).

28 S. E. Kahn, R. L. Hull, and K. M. Utzschneider, *Nature* **444**, (2006).

29 A. Cesar-Razquin, B. Snijder, T. Frappier-Brinton, R. Isserlin, G. Gyimesi, X. Bai, R. A. Reithmeier, D. Hepworth, M. A. Hediger, A. M. Edwards, and G. Superti-Furga, *Cell* **162**, (2015).

30 L. J. Baier, P. A. Permana, X. Yang, R. E. Pratley, R. L. Hanson, G. Q. Shen, D. Mott, W. C. Knowler, N. J. Cox, Y. Horikawa, N. Oda, G. I. Bell, and C. Bogardus, *J Clin Invest* **106**, (2000).

31 F. Shi, Z. Xu, H. Chen, X. Wang, J. Cui, P. Zhang, and X. Xie, *Monoclon Antib Immunodiagn Immunother* **33**, (2014).

32 K. T. Yip, X. Y. Zhong, N. Seibel, S. Putz, J. Autzen, R. Gasper, E. Hofmann, J. Scherkenbeck, and R. Stoll, *Sci Rep* **6**, (2016).

33 S. A. Ainger, X. L. Yong, S. S. Wong, D. Skalamera, B. Gabrielli, J. H. Leonard, and R. A. Sturm, *Exp Dermatol* **23**, (2014).

34 C. Ma, and A. W. Armstrong, *J Dermatolog Treat* **25**, (2014).

NETWORK MAP OF ADVERSE HEALTH EFFECTS AMONG VICTIMS OF INTIMATE PARTNER VIOLENCE

KATHLEEN WHITING

Neuroscience Program, Uniformed Services University, 4301 Jones Bridge Rd,
Bethesda, Maryland 20814, USA
Email: kathleen.whiting@usuhs.edu

LARRY Y. LIU

Center of Proteomics and Bioinformatics, Case Western Reserve University, 10900 Euclid Ave,
Cleveland, Ohio 44106, USA
Email: lyl14@case.edu

MEHMET KOYUTÜRK

Department of Electrical Engineering & Computer Science, Case Western Reserve University, 10900 Euclid Ave,
Cleveland, Ohio 44106, USA
Email: mxk331@case.edu

GÜNNUR KARAKURT

Department of Psychiatry, Case Western Reserve University, 10900 Euclid Ave,
Cleveland, Ohio 44106, USA
Email: gkk6@case.edu

Intimate partner violence (IPV) is a serious problem with devastating health consequences. Screening procedures may overlook relationships between IPV and negative health effects. To identify IPV-associated women's health issues, we mined national, aggregated de-identified electronic health record data and compared female health issues of domestic abuse (DA) versus non-DA records, identifying terms significantly more frequent for the DA group. After coding these terms into 28 broad categories, we developed a network map to determine strength of relationships between categories in the context of DA, finding that acute conditions are strongly connected to cardiovascular, gastrointestinal, gynecological, and neurological conditions among victims.

1. Introduction

Domestic abuse is a rampant problem across the globe, contributing to severe economic, health related, and societal costs. The consequences of intimate partner violence (IPV) are devastating and systemic. In 2010, the National Intimate Partner and Sexual Violence Survey found that approximately 30% of women experience physical violence from an intimate partner during their lifetime, with 25% experiencing severe physical violence such as being slammed, hit, or beaten.[1] Although victims of IPV are not exclusively female, women are more likely than men to be the victim and sustain serious physical injury.[1,2]

IPV has been shown to cause numerous adverse health effects, ranging from minor injuries to serious disability and death.[2-4] Physical assault (including sexual violence) is associated with psychological distress such as anxiety, depression, and suicidal ideation,[5] sexually transmitted infections including HIV,[3,6] gynecological problems like pelvic inflammatory disease,[7] and unintended pregnancy and complications relating to the mother's and newborn's health.[3,8-10] Researchers have identified many long-term effects of IPV, finding evidence that victims of violence are more prone than the general populous to suffer from mental health and substance abuse disorders, gastrointestinal problems, chronic pain and physical ailments, and various neurological symptoms.[5,11] This information has not prompted further research into how professionals can prevent and treat the IPV epidemic. Holistic approaches incorporating comprehensive treatment of both physical and emotional ailments have received little attention.

Self-report data indicates that female victims of violence have poorer overall health than female victims of non-violent crimes (and women in general), presenting troubling physical symptoms like tachycardia, tension headaches, menstruation related issues, stomach problems, or skin disorders.[12] Intimate partner violence often involves episodes of physical and sexual violence. No doubt this is a contributing factor to poor victim health, and likely increases the use of healthcare services by victims of violence. For example, sexual assault victims are more likely to seek physical and mental health care within the first six months of the attack, with services increasing 15-24% during the first twelve months alone. Naturally, the corollary of this is a higher cost of health care and treatment for victims than non-victims. Emergency room records indicate un-witnessed episodes of head, neck, and facial injuries are significant markers of IPV,[13] and traumatic brain injury may be more prevalent in this population than previously suspected.[14] Unfortunately, physicians often overlook or misattribute problems associated with violence, which can result in prolonging victims' pain and wasting patient and provider resources. Proper screening and treatment of IPV is critical to ensure that victims of violence receive the necessary care and support for recovery.

While the effects of IPV are known to be serious and diverse, knowledge of specific health effects and their relation to IPV is still limited. In this study, we utilized electronic health records (EHR) to identify frequently occurring symptoms among IPV victims. Our approach is motivated by the notion that EHR data provide valuable information from health care providers that may not be obtained through self-report data. Furthermore, both self-report data and physicians' records are difficult to obtain in large amounts due to topic sensitivity. For these reasons, investigators struggle to compile available symptom data into comprehensive and systematic reviews. The consequences of violence on human health are elusive and complex, and therefore utilization of large-scale data can be useful in identifying correlates that are overlooked by other research. Here, we take a first step toward utilizing EHR data to characterize adverse health effects co-occurring with IPV, identifying statistical associations between IPV and other symptoms and determining the strength of these relationships. It is important to note that our analysis does not target any symptom in particular; rather we mine the entire EHR data (1999 through our original data query point 5/8/14) and test the association of all reported symptoms to identify those statistically significant.

In a previous study,[15] we accessed and analyzed national EHR data through the *Explorys* platform (Explorys Inc., an IBM company), specifically utilizing the "Explorys Enterprise Performance Management (EPM): Explore" web application to identify diseases which seem to be more prevalent among victims of IPV than the general US population. *Explorys* is comprised of EHR, EMR, insurance claims, and billing data sources. A variety of national data sources contribute data to the platform, including affiliated providers, electronic medical systems, health care plans, and care settings. Over twenty major integrated healthcare systems provide data to *Explorys*, bringing together patient information from across America. Over 300,000 providers participate, gathering more than 315 billion clinical, operational, and financial data elements from approximately 50 million unique patients. Data is pooled from clinical EMRs, healthcare system outgoing bills, and adjudicated payer claims. Researchers from a wide range of disciplines use this compiled data to identify patterns and trends in diseases, treatments, and outcomes.[16]

We hoped that our analysis of the data we obtained through the *Explorys* platform would help us differentiate between those health problems which result directly from acute violence and violence related physical injuries, and those health problems which are chronic or persistent and result from multiple non-violent causes. After identifying the diseases occurring significantly more frequently among victims of IPV, we categorized the diseases into 28 broad categories, and found that IPV is predominantly associated with four types of health problems: acute; chronic; gynecological; and mental/behavior health. The results further supported our suspicions that IPV is a systematic problem with multifaceted interactions across a wide range of health issues.

To develop a better understanding of how IPV is related to negative health effects, it is potentially useful to determine the interactions and relationships between symptom categories. Analyzing these relationships may help us discover what physiological systems are more closely associated with experiencing severe consequences of IPV, and could lead to future research into the effects of IPV on the body. For this study, we decided to perform a data-driven analysis and network mapping of the same significantly occurring diseases identified previously to reveal how these terms interact. We chose network mapping because it analyzes the structural relationships and patterns within a network of 'nodes', providing a visual representation of the strength of these relationships. In this case, the nodes are each symptom category, and the connections or 'edges' between these categories indicate how frequently those given categories appear together in our coded symptoms. Through this analysis we specifically hope to identify the strength of connections between different disease categories, in an effort to investigate how these associations may be related to each other. Through this analysis we can explore the many ways IPV affects the overall health of victims.

2. Methods

2.1. *Identification of Terms Prevalent among Domestic Abuse Victims*

The flow chart for the methodology implemented in this study is shown in Figure 1. A complete and detailed description of the data acquisition performed for this study can be found in a previously published manuscript.[15] Here, we provide a brief summary.

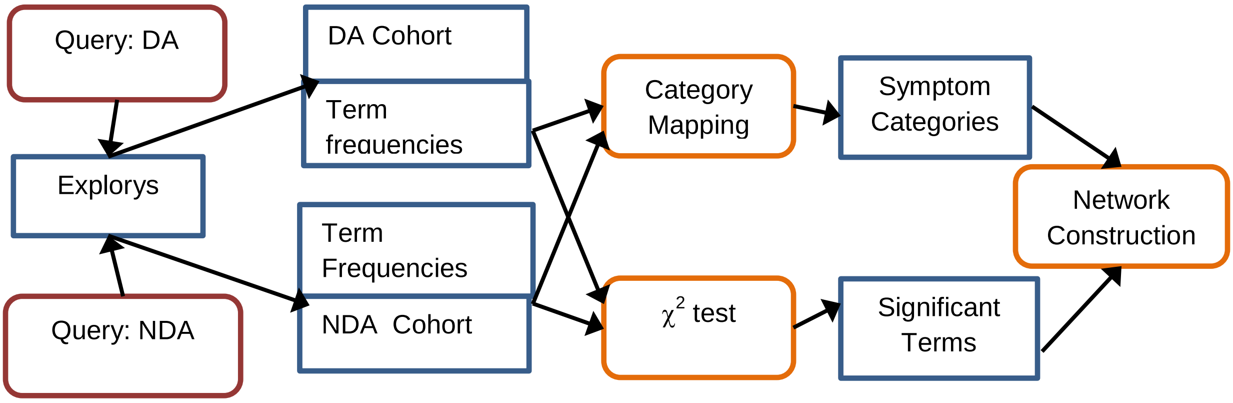


Fig 1. Flowchart for data acquisition, querying, statistical analysis, and network construction. DA: Domestic Abuse. NDA: Non-Domestic Abuse

We obtained data from a population of almost 15 million patients who were adult females aged 18-65 seen in multiple different healthcare systems across the United States with unique EHR from 1999 to the present (5/8/14: original query date). These data were normalized and classified using common ontologies, searchable through the HIPAA-enabled, de-identified "Explorys Enterprise Performance Management: Explore" web application. Using SNOMED clinical terms built into *Explorys*, our search query identified 5870 records (DA cohort) of IPV victims (these records were retrieved by searching for the finding 'domestic abuse', a code option utilized by health professionals for EHR), compared to 14,315,140 records (NDA cohort) of patients who did not have any indication of IPV victimization in their EHR. Racial and age distribution for the DA and NDA cohorts are shown in Figure 2.

It is important to note that in order to protect patient privacy, data is accessible only as frequencies across the cohorts defined by these queries. Similarly, demographic information is available as summaries. For this reason, sophisticated data mining techniques such as association rule mining are not directly applicable. Here, we base our analysis on the comparison of frequencies. Of note, African Americans make up a greater proportion of the DA group than NDA, and while NDA records are relatively evenly distributed across age groups, DA records show higher relative frequencies for ages 25-44.

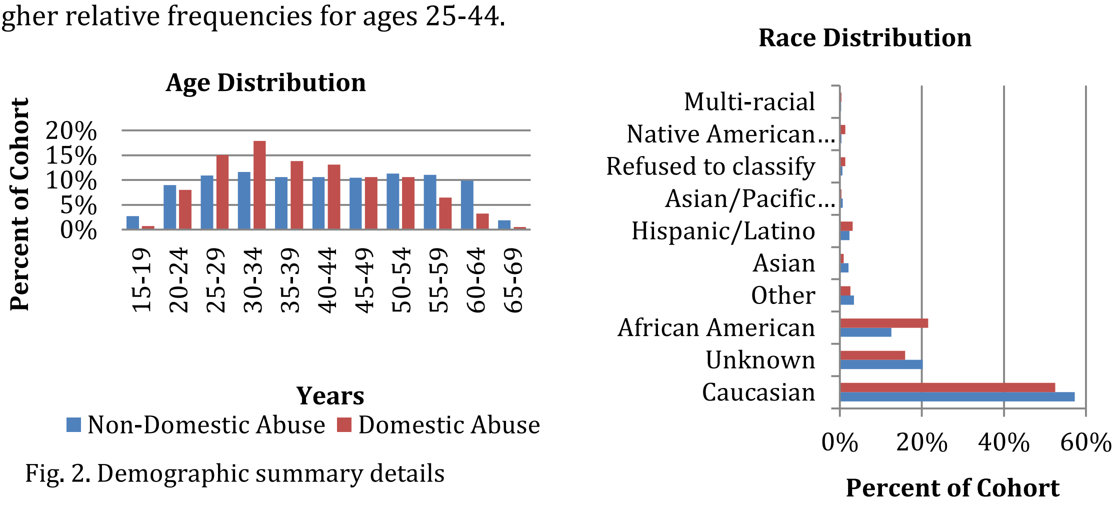


Fig. 2. Demographic summary details

Within the IPV identified records, we found 3458 symptom/diagnosis terms possibly associated with domestic abuse (i.e. the aggregated records contained 3458 medically coded symptoms, diagnoses, and findings – also referred to as "terms" in the following discussion). With a view to identifying conditions that are prevalent among victims of domestic abuse, we compared the frequency of each term in the DA cohort with its frequency in the NDA cohort. For each term, we used χ^2-test to assess the significance of its frequency in the DA cohort with respect to its frequency in the NDA cohort. To adjust for multiple hypothesis testing, we used Bonferroni correction (3458 tests were performed). This analysis suggested that 2430 of the identified terms were significantly more prevalent among IPV victims ($p < 0.05$). Two independent researchers used medical dictionaries to manually code these symptoms into broader, more general categories with high inter-rater reliability, four main classes emerged: chronic symptoms and disorders, acute injuries, mental and behavioral issues, and gynecological problems.

2.2. *Network Construction*

To provide a compact and visually comprehensive view of the diagnosis terms that were significantly more frequent in patients with the finding of "domestic abuse" (DA group), we created a network of diagnosis categories.

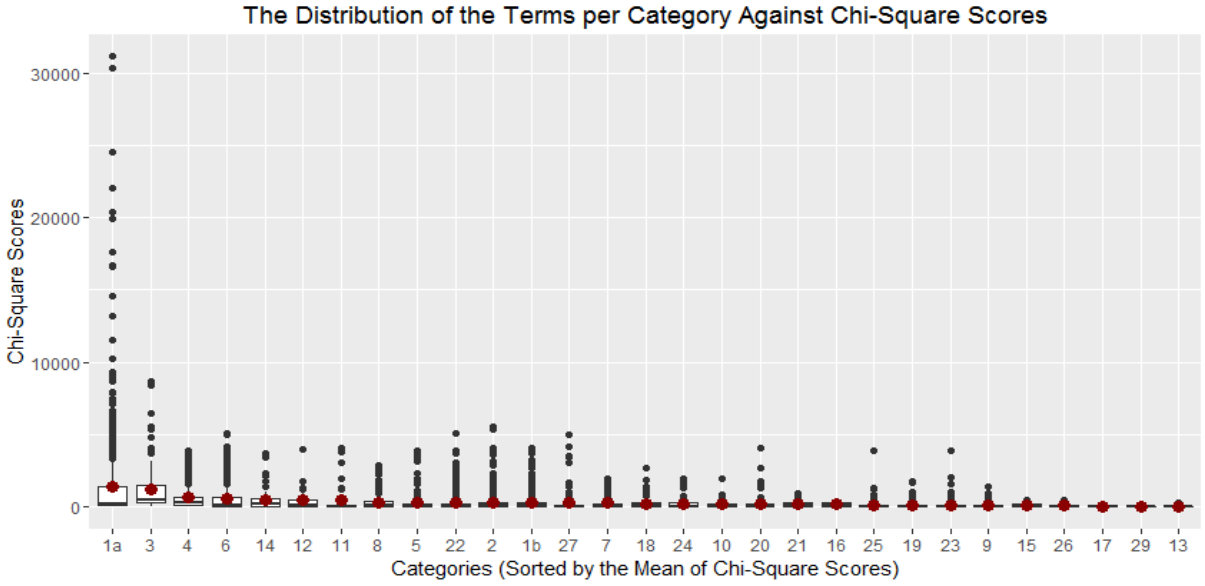


Fig. 3. Distribution of χ^2-statistics among terms assigned to each category. The distributions are shown by box plots, the mean of each distribution is shown by a red. The categories are sorted according to the mean χ^2 statistic

We first categorized the diagnosis terms by assigning each term to 28 specific categories. In this classification, assignment of a term to more than one category was allowed. Subsequently, we selected the 2429 terms that were significantly more frequent in the DA group ($p< 0.05$ based on the Chi-Square Test). We then counted the frequency of each category among these 2429 terms. The distribution of χ^2-statistics for the terms assigned to each category is shown in Figure 3.

	Disease Classification Category	Percent (%)
1b	Acute Condition	34.1%
1a	Acute Injury	23.1%
2	Chronic	16.2%
6	Disorders	15.6%
19	Cardiovascular	8.6%
8	Pregnancy Related	7.5%
7	Gynecological	7.5%
22	Musculoskeletal	6.2%
4	Mental Health	5.9%
18	Gastrointestinal	5.6%
9	Allergy	5.0%
3	Substance Abuse	4.8%
5	Other	4.6%
20	Nervous system	4.5%
27	Skin related (not burns)	4.0%
21	Respiratory	3.9%
23	Eyes, Ears, Nose & Throat	3.8%
24	Excretory	3.1%
14	Personal History	2.2%
11	Congenital/Hereditary	1.6%
25	Endocrine	1.6%
13	Neoplasm	1.3%
26	Immune System	1.3%
12	Nutrition	1.2%
10	Procedure	0.8%
16	Neuropathy	0.8%
15	Family History	0.7%
17	Diabetes	0.6%

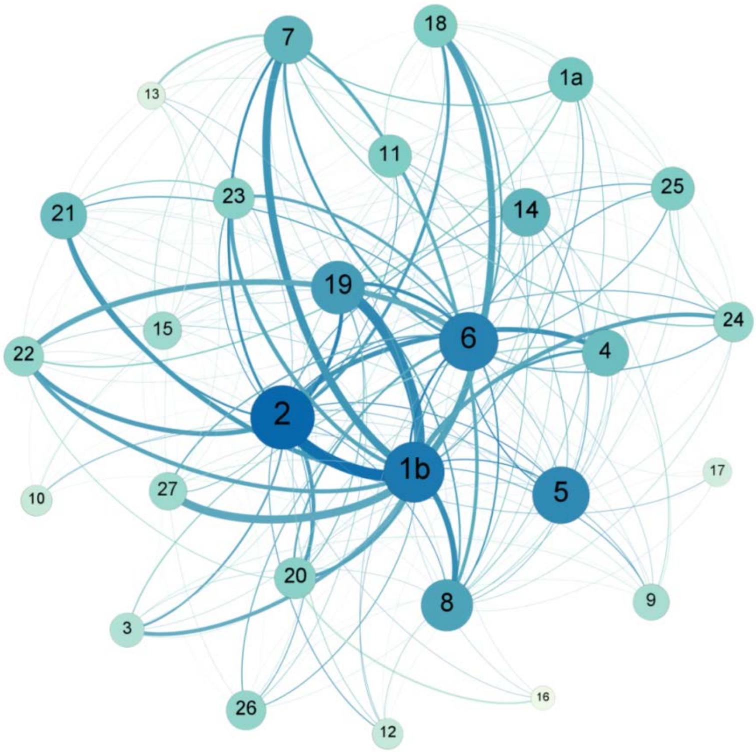


Fig. 4. Network Map of the 28 categories made up of 2429 symptom terms found to be significantly more prevalent among victims of IPV than the general population. Larger nodes indicate higher coding frequency of the category by itself, while thicker edges reveal the strength of relationship between two nodes by signifying the frequency of any two categories being coded together on a single symptom term. Darker nodes appear more frequently among the pairs than lighter nodes. Percentages of total diagnoses found to be significantly frequent in patients with a DA finding for each individual symptom category also shown.

To assess the co-occurrence of each pair of categories in the DA group, we counted the number of terms among these 2429 terms that are assigned both categories. Note that this co-occurrence frequency is not to be confused with co-morbidity of diagnoses; this number rather reflects the likelihood that a diagnosis significantly more frequent in patients with DA will belong to both categories. In total, 208 pairs of categories were assigned together to at least one of the 2429 terms.

After counting frequencies of categories and pairs of categories among the terms that are significantly frequent in the DA group, we visualized the frequencies as a network using GePhi 0.9.0 Beta (Fig. 4). In the figure, the size of each node represents the frequency of the respective category among the terms that are significantly frequent in the DA group. The thickness of each edge represents the number of terms that are assigned both of the respective categories.

3. Results

The results of the data-driven analysis and network mapping are illustrated in Figure 4. The Acute Conditions (1b) and Chronic (2) categories appear to be the most significant in our network map, showing the greatest frequency of occurrence among the coded terms. Chronic (2) exhibits strong connections to Acute Conditions (1b), and fair connections to Mental Health (4), Cardiovascular (19), Nervous System (20), and Musculoskeletal (22). It should be noted that the strong connection between chronic and acute conditions is due to ambiguity of the symptom terms, resulting in the possibility of a term being coded as both chronic and acute because it could be either, depending on the patient's situation. Acute Conditions (1b) has strong connections to Chronic (2) and Cardiovascular (19), with more moderate connections to Gynecological (7), Gastrointestinal (18), Skin Related (27), and Pregnancy Related (8). It also has fair connections to Substance Abuse (3), Nervous System (20), Respiratory (21), Musculoskeletal (22), Eyes, Ears, Nose & Throat (23), and Excretory (24).

Disorders (6) shows some significance in coded frequency, with a moderate connection to the Musculoskeletal (22) category. The Cardiovascular (19), Gynecological (7), Pregnancy Related (8), and Gastrointestinal (18) nodes may also be somewhat significant, but their primary connection is with Acute Conditions (1b), which is itself a significant node in the network overall (Cardiovascular is also fairly connected to Chronic (2), another independently significant node).

Interestingly, although the Acute Injures (1a) node indicates relatively significant frequency of coding, this category is mostly isolated, demonstrating only weak connections.

4. Discussion

4.1. *General*

The results of our analysis and network mapping are fairly consistent with current knowledge of IPV. We found that chronic and acute conditions as well as acute injuries were frequently coded to the symptoms that are more prevalent among victims of IPV. We also found that the categories showing the most individual frequency (chronic and acute conditions) shared strong connections with physiological systems that have been shown to be impaired at a higher rate among IPV victims, including gynecological and pregnancy related issues, as well as gastrointestinal, cardiovascular, and neurological symptoms.[17-19] It is not surprising that 'chronic' and 'acute conditions' had the most significant frequency of coding, since most ailments that require medical attention are either established diseases or emergency issues. 'Acute conditions' shows many strong connections, because most acute conditions would be coded with whatever body system(s) they were related to. 'Chronic' is similarly highly connected, for the same reason. These two categories show a strong connection to each other because the symptoms were often ambiguous, and coded as both chronic and acute because there was not enough information in the symptom term to differentiate it between the categories.

Nodes that represent 'gynecological' and 'pregnancy related' symptoms appear to be fairly significant, which is expected when considering the nature of IPV. Physical and sexual abuse from IPV can be very damaging to the body,[17,20] resulting in trauma, infection, and the contraction of sexually transmitted infection.[6] Victims are at greater risk of experiencing sexual coercion from an intimate partner as well as birth control sabotage, and often fear talking to their partner about pregnancy prevention.[6] Studies have also shown increased risk to mothers' and newborns' health when IPV is experienced during pregnancy, such as miscarriage and low birth weight.[8,10] Pregnancy itself can also be a risk factor for IPV.[21]

We were surprised to see that the node that represents 'mental health' related symptoms was not a significant "hub" in the network. The node itself is significant, i.e., it appears moderately prevalent in the general frequency count, but lacks strong connections. This independence is explained by the fact that most symptoms labeled as 'mental health' would be unlikely to fall into other categories except for 'chronic', since many mental health issues happen to be chronic in nature but generally would not directly interact with other physiological systems.

Stress may be an important factor in the patterns we identified in the network map and analysis. There is a broad research literature describing the interactions between IPV and stress,[17,22,23] as well as the effects of stress on the body.[24-26] IPV causes an increase in cortisol, one of the body's stress hormones, which in turn might cause detrimental impacts on the victim's immune system. This can manifest in a variety of ways, but often affects gastrointestinal and circulatory system functioning. Stress resulting from IPV may also seriously increase the risk of a negative event during and following pregnancy.[27,28] Interestingly, the nodes that represent these categories showed significant frequency in the network of IPV victim health symptoms. It is difficult to verify if stress is the underlying factor in the higher significance and connection of these categories, but it may be possible in future studies to incorporate cortisol-level measurements in the search query. Many of the significant categories in our network map showed strong connections to 'chronic' and 'acute conditions', which only further demonstrates the severity of the negative health effects associated with IPV.

We cannot accurately assess whether this network map reveals subtle patterns or cycles, because our data is a compilation of records that incorporate data without the dimension of time. If we could find a way to apply temporal filters, we might be able to identify a progression of health events common among victims of IPV that would further illustrate exactly how IPV leads to negative health over the victim's lifetime. This will enable health care practices and professionals to more accurately identify, assess, and treat IPV and related illnesses, ultimately utilizing knowledge of how IPV impacts health to improve treatment decision making processes. Analyzing this data will help explore how EHR can be utilized for research. Our findings may demonstrate that it is possible to improve standard screening procedures and treatment plans for victims of violence as well as patients in other circumstances, simply by examining electronic health records. The significant correlations that can be found through this method provide valuable information for both clinical and research applications.

4.2. *Limitations*

Although studies have shown that approximately 1 in every 4 women will experience IPV at some point during their lifetime,[1] the data collected by *Explorys* does not reflect this observation. This is likely due to a variety of factors, and may be most influenced by the underreporting and inadequate screening of IPV victims. Although our query returned NDA records from all 50 states, as well as Puerto Rico, Guam, and APO/FPO (military bases), it returned DA records from only a dozen location categories. The relative distribution of records across these location categories for both DA and NDA cohorts are shown in Figure 5. It would be interesting to examine the laws and regulations of the states that returned EHR for the DA group. It is possible that local regulations influence the likelihood of domestic abuse being screened and recorded by medical professionals.

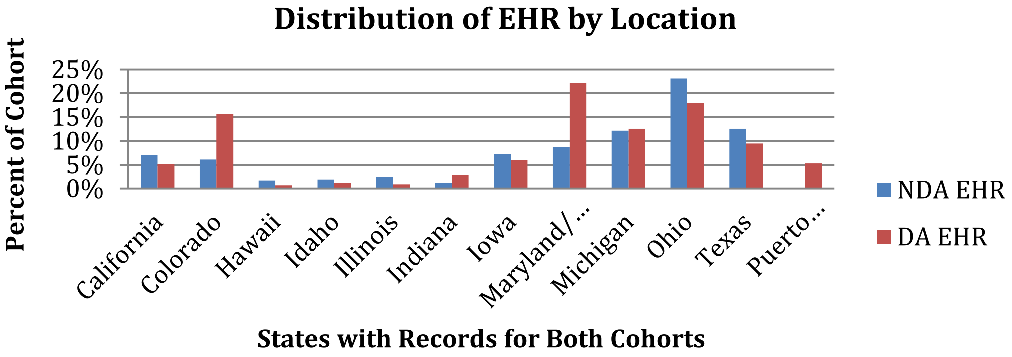


Fig. 5. . Distribution of location categories for the domestic abuse (DA) and non-domestic abuse (NDA) cohorts.

It is interesting to note that the distribution of EHR for the NDA cohort does not correspond with the distribution of the US population as measured by the 2010 US Census Bureau (comparison not shown here, but NDA EHR showed disproportionately high distribution in Texas, Ohio, Michigan, Maryland/DC, Iowa, Idaho, Hawaii, and Colorado compared to relative census population distributions). This difference may reflect the sampling of medical facilities providing data to *Explorys*. These trends are worth investigating to better understand how inherent confounds of the data may influence query results. Many of the records pertaining to this subset of the population may not have been captured by our queries, and thus potentially were not included in our network map and data analysis. This means that it is very likely that records in our NDA group actually belong in the DA group, which would seem to muddy the analysis. However, we feel that due to this underreporting, our DA group likely represents the most severe cases of domestic abuse, and thus highlights the most obvious symptom connections. Further analysis could help us tease out more subtle victim symptom characteristics. This information could then be used to develop targeted queries in the future that may help us to identify high-risk IPV victims through related EHR diagnoses, even if domestic abuse is not listed as a finding. While we were able to extract some demographic characteristics of the cohorts (Fig. 2), the nature of the data does not permit us to match demographics to specific records. However, as Fig. 5 illustrates, there are serious gaps in EHR data for the DA group, and it is not possible to identify confounds resulting from demographics with such limited data.

Domestic abuse is not always noted in a patient's medical records, either because the patient does not reveal that she is a victim, or because medical professionals overlook recording this detail. Research has demonstrated that primary care IPV screening is inadequate and needs attention.[29] Domestic abuse is not considered a 'diagnosis' or 'condition', but is rather described in the *Explorys* database as a 'finding'. This is a reflection of how the medical community labels IPV, and explains why notating this 'finding' may not be a priority when updating a patient's medical records. Since this study most likely captured records with the most severe cases of IPV, where evidence of the abuse was obvious, the current network map may not reflect the more subtle connections associated with less severe instances of IPV. However, the results of our network map and analysis demonstrate how extensive the consequences of IPV may be on victims' health, and further illustrate the vital importance of thoroughly screening for IPV and accurately noting patient findings by members of the medical community. It may be possible to utilize this data to develop more accurate screening procedures in the future.

Utilizing EHR data is challenging, and the currently available query systems leave gaps in data quality. Data is noisy, and we lack the ability to control for a myriad of confounds. Current records allow for the possibility of patients being counted more than once when determining frequencies, and longitudinal data is missing. However, due to the necessity to maintain the strictest levels of privacy, we cannot track specific (though still de-identified) patient records to see how health changes over time. We are exploring how to utilize other tools in *Explorys* and other query and analysis techniques to capture longitudinal data. Examining the changes in symptom presentation in relation to the first appearance of DA on a patient's medical record is a key step to understanding the etiology and health consequences of violence victimization more fully. This could also be an instrumental step in implementing effective risk assessments. However, even at this point, the knowledge gained from the analysis of EHR data can still lead to vast improvements in health care and policy development, and improved queries may improve the quality of data. The techniques demonstrated in this study have implications not only for the care of intimate partner violence victims, but for the health of the entire population as a whole.

5. Conclusion

EHR data is a vital resource in advancing the knowledge of health care professionals. By analyzing the data we can create networks that show how different symptom and disease categories are related to each other, revealing associations which may indicate deeper root causes for deteriorating health. In this study, we were able to examine what health factors are associated with IPV, and how these factors interact. This gives us a more complete and compelling picture of the negative health effects of IPV. With further research it may be possible to develop improved methods and diagnostic tools for successful intervention and treatment, improving victims' quality of life throughout their lives.

Analysis of EHR data gives health providers the information to improve quality of service, especially for victims of IPV. However, it is so important for screening procedures to improve, so that victims are accurately identified and given appropriate medical care. Our network mapping and data analysis demonstrate only a fraction of the far-reaching health consequences of IPV, which cannot be ignored from a medical perspective. We know that many of the victims of IPV were not represented in our DA data set, because they haven't been identified as such in their medical records. It is absolutely imperative that we push to improve screening so that these devastating health effects can be mitigated and prevented. The data from our analysis may help with future research into how we can better identify victims who hesitate to come forward by identifying the tell-tale signs and relationships of their symptoms and conditions. It is clear that mining EHR will reveal many associations between previously independent conditions. Our future research will replicate these analysis techniques with independent datasets to confirm the efficacy of these methods. Doctors and health care providers can use this information to improve the prescription of effective treatment preventions, and identify trends across populations. If we can use this information to develop more effective screening tools and treatments, we will drastically increase the quality of life and healthcare experienced by victims of IPV, and through this the wellbeing of society as a whole.

6. Acknowledgements

This publication was made possible in part by the Clinical and Translational Science Collaborative of Cleveland, NIH/NCRR CTSA KL2TR000440 from the National Center for Advancing Translational Sciences (NCATS) component of the National Institutes of Health and NIH roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References

1. M. C. Black, K. C. Basile, M. J. Breiding, S. G. Smith, M. L. Walters and M. T. Merrick, National intimate partner and sexual violence survey 2010 summary report and sexual violence survey (2011).

2. American College of Obstetricians and Gynecologists. *Obstet Gynecol,* 119, 412 (2012).

3. J. Corrigan, M. Wolfe, W. Mysiv, R. Jackson and J. Bogner, *Am J Obstet Gynecol*, 188, S71 (2003).

4. E. Lawrence, A. Oringo and R. Brock, *Partner Abuse,* 3 (2012).

5. G. Karakurt, D. Smith and J. Whiting, *J Fam Violence,* 29 (2014).

6. G. Wingood, R. DiClemente, D. McCree, K. Harrington and S. Davies, *Pediatr,* 107, E72 (2001).

7. E. Letourneau, M. Holmes and J. Chasedunn-Roark, *Women's Health Issues,* 9, 115 (1999).

8. D. El Kady, W. Gilbert, G. Xing and L. Smith, *Obstet Gynecol,* 105, 357 (2005).

9. J. Hathaway, L. Mucci, J. Silverman, D. Brooks, R. Mathews and C. Pavlos, *Am J Prev Med,* 19, 302 (2000).

A. Huth-Bocks, A. Levendosky and G. Bogat, *Violence Vict*, 17, 169 (2002).

10. H. Nelson, C. Bougatsos and I. Blazina, *Ann Int Med, 156*, 796 (2012).

11. J. Campbell, *The Lancet, 359,* 1331 (2002).

12. V. Wu, H. Huff and M. Bhandari, *Trauma Violence Abuse, 11*, 71 (2010).

13. L. Kwako, N. Glass, J. Campbell, K. Melvin, T. Barr and J. Gill, *Trauma Violence Abuse, 12*, 115(2011).

14. G. Karakurt, V. Patel, K. Whiting and M. Koyuturk, *J Fam Violence*, in print, (2016).

15. D.C. Kaelber, W. Foster, J. Gilder, T. E. Love and A. K. Jain, *J Am Med Inf Assoc, 19,* 965 (2012).

16. J. Campbell, A. S. Jones, J. Dienemann, J. Kub, J. Schollenberger, P. O'Campo, A. C. Gielen and C. Wynne, *Arch Int Med, 162,* 1157 (2002).

17. D. J. Sheridan and K. R. Nash, *Trauma Violence Abuse, 8*, 281 (2007).

18. S. Sprague, K. Madden, S. Dosanjh, K. Godin, J. C. Goslings, E. H. Schemitsch and M. Bhandari, *BMC Musculo Disorders, 14* (2013).

19. M. L. Paras, M. H. Murad, L. P. Chen, E. N. Goranson, A. L. Sattler, K. M. Colbenson, M. B. Elamin, R. J. Seime, L. J. Prokop and A. Zirakzadeh, JAMA, 302, 550 (2009).

20. T. L. Taillieu and D. A. Brownridge, *Aggress Violent Behav, 15*, 14 (2010).

21. M. E. Feinberg, D. E. Jones, D. A. Granger and D. Bontempo, *Aggress Behav, 37*, 492 (2011).

22. S. S. Inslicht, C. R. Marmar, T. C. Neylan, T. J. Metzler, S. L. Hart, C. Otte, S. E. McCaslin, G. L. Larkin, K. B. Hyman and A. Baum, *Psychoneuroendocrinology, 31*, 825 (2006).

23. M. Moreno-Smith, S. K. Lutgendorf and A. K. Sood, *Future Oncol, 6*, 1863 (2010).

24. M. Roest, E. J. Martens, P. de Jonge and J. Denollet, *J Am Coll Cardiol, 56*, 38 (2010).

25. J. Shen, Y. E. Avivi, J. F. Todaro, A. Spiro, J. P. Laurenceau, K. D. Ward and R. Niaura, *J Am Coll Cardiol, 51*, 113 (2008).

26. A. Taylor, N. B. Guterman, S. J. Lee and P. J. Rathouz, *Am J Pub Health, 99*, 175 (2009).

27. W. P. Witt, E. R. Cheng, L. E. Wisk, K. Litzelman, D. Chatterjee, K. Mandell and F. Wakeel, *Am J Pub Health, 104*, S81 (2014).

28. P. Tavrow, B. E. Bloom and M. H. Withers, *Violence Against Women*, in print (2016).

DISCOVERY OF FUNCTIONAL AND DISEASE PATHWAYS BY COMMUNITY DETECTION IN PROTEIN-PROTEIN INTERACTION NETWORKS

STEPHEN J. WILSON

Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza
Houston, Texas 77030, USA
Email: sw5@bcm.edu

ANGELA D. WILKINS

Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza
Houston, Texas 77030, USA
Email: aw11@bcm.edu

CHIH-HSU LIN

Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, One Baylor Plaza
Houston, Texas 77030, USA
Email: Chih-Hsu.Lin@bcm.edu

RHONALD C. LUA

Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza
Houston, Texas 77030, USA
Email: lua@bcm.edu

OLIVIER LICHTARGE

Departments of Molecular and Human Genetics, Structural and Computational Biology and Molecular Biophysics, Biochemistry and Molecular Biology, and Pharmacology, Baylor College of Medicine, One Baylor Plaza
Houston, Texas 77030, USA
Email: lichtarg@bcm.edu

Advances in cellular, molecular, and disease biology depend on the comprehensive characterization of gene interactions and pathways. Traditionally, these pathways are curated manually, limiting their efficient annotation and, potentially, reinforcing field-specific bias. Here, in order to test objective and automated identification of functionally cooperative genes, we compared a novel algorithm with three established methods to search for communities within gene interaction networks. Communities identified by the novel approach and by one of the established method overlapped significantly ($q < 0.1$) with control pathways. With respect to disease, these communities were biased to genes with pathogenic variants in ClinVar ($p << 0.01$), and often genes from the same community were co-expressed, including in breast cancers. The interesting subset of novel communities, defined by poor overlap to control pathways also contained co-expressed genes, consistent with a possible functional role. This work shows that community detection based on topological features of networks suggests new, biologically meaningful groupings of genes that, in turn, point to health and disease relevant hypotheses.

1. Introduction

How genes and proteins interact with each other is the basis of molecular biology and disease pathogenesis[1,2]. These functional interactions, which biologists place into pathways, have been characterized through hypothesis-driven experiments and then manually defined in the past[3,4]. This is necessarily knowledge intensive and painstaking, and it stands in sharp contrast to the massive amount of new gene interaction data from high-throughput experiments. Continued reliance on manual recognition of pathways may limit the overall capacity to characterize gene behavior, and potentially focus on already well-known sets of gene interactions. With at least 100,000 interactome hubs in humans, the number of potential interactions to annotate are in the billions[5]. Yet, the current estimate of interactions from the broadly used and expertly curated STRING database[6] that focus solely on proteins are in the millions. This large discrepancy suggests many unrecognized, or "dark," associations and pathways are simply missing.

In order to take a data-driven approach to annotate and detect novel biological pathways, clusters in biological networks were defined based on topological features to isolate functional and disease pathways[5,7]. One topological feature that has been extensively applied in social network analysis[8-10], but has not yet seen widespread use in biology, is community structure[11,12].

Communities are groups of nodes (i.e. proteins) that are more connected to each other than to anything else in a network[8,13]. Often these groups of nodes correspond to a common process, purpose, or function[5,9]. Therefore, it is reasonable to hypothesize that determining communities on biological networks may shed new light on groupings of genes with common biological function or features. Past efforts[13,14] were useful but did not comprehensively test various algorithms in functional and disease contexts. Given appropriate algorithms, community detection has the potential to automatically expand biological pathways, determine novel pathways, and perhaps even predict gene-disease associations.

This study sought to detect communities on a protein-protein interaction network and to evaluate their number and size against existing references. Several methods can evaluate performance in terms of the number and size of the overlap between communities and known control pathways. Moreover, beyond reference pathways, disease data can directly demonstrate the applicability of communities to formulate new and clinically relevant biomedical hypotheses.

2. Results

2.1. *Determining putative biological pathways*

In order to automatically determine putative biological pathways, several possible community detection methods exist. Clauset-Newman-Moore (CNM)[8] and Louvain[10] are well-established and extensively applied algorithms with more than 3000 citations each. BIGCLAM[15] is a more recent alternative that searches for densely overlapping, hierarchically nested communities in an orthogonal approach. Each of these approaches was tested on a STRING protein-protein interaction network[16], limited to high-quality direct biological associations. The communities that were obtained could then be compared to gold standard set of curated biological pathways, such as Reactome[17] and Canonical pathways from the GSEA tool[18], and, for disease pathways, DisGeNET[19].

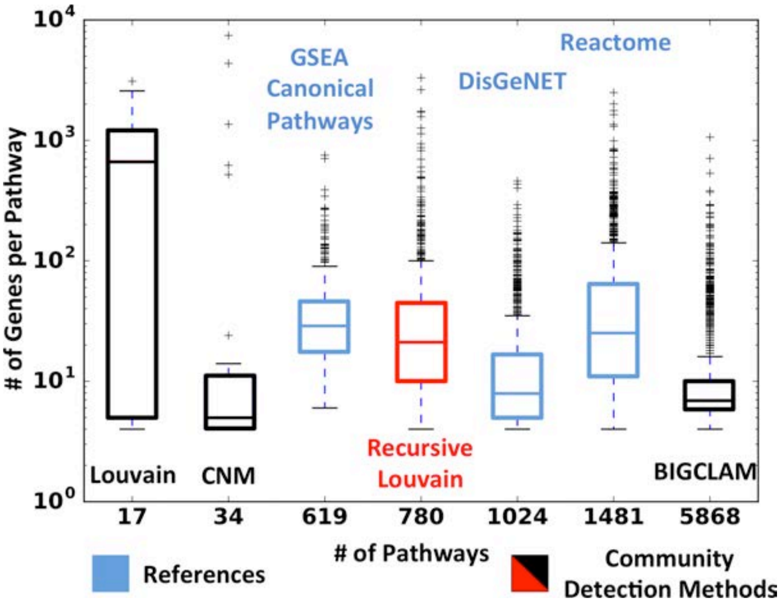


Figure 1: Community Algorithms Detect Variable Numbers and Sizes of Pathways. Recursive Louvain (Red) is a novel community detection method that detects a similar number of gene groups to the references with approximately the same number of genes per gene group given STRING 9.1 protein-protein interaction network.

A first assessment of performance was the granularity of the communities. That is, we compared the number of gene groups and the number of genes in each group in order to determine whether the communities resemble the references. CNM and Louvain community detection found an order of magnitude fewer groupings than the smallest reference set, and BIGCLAM detects five times more groups than the largest reference set (Figure 1). This is not surprising given that the methods were designed for social network analysis. Combined with different numbers of genes per group, these algorithms appear to poorly represent the reference pathways as defined by biologists.

To address this problem, we then introduced a novel community detection algorithm we called, Recursive Louvain (RL). RL applies the Louvain algorithm but iterates on the resulting communities so as to break them down stepwise into smaller and smaller groups until reaching a majority of right-sized communities, an idea that was in fact discussed in the original Louvain community detection paper[10]. In this way, RL generated communities that matched more closely the size of the control pathways (Figure 1, red).

2.2. *Assessing the biological relevance of communities*

Next, when comparing communities to the reference sets, careful consideration of what constitutes a pathway was necessary. First, we removed overly small reference pathways and communities (size ≤ 3 genes) to better focus on significant gene groupings. Additionally, pathways often share many genes, and in the extreme, they can share all but one gene. To avoid the over-counting of a pathway, or community, any with more than 90% of genes in common were combined. Finally, four different metrics were selected to gauge success. Jaccard similarity measures the similarity of a community to a reference by looking at the size of the intersection relative to the union of the genes; the modified Jaccard metric does not punish a community for being larger than the reference; the hypergeometric test measures the likelihood of getting an overlap between a reference and a community given all genes in a given community set; and the F_1 score measures the ability to recover an overlap (see methods for the mathematical details).

To test if communities represent biological information from functional and disease pathways, we compared each community to each reference pathway. This comparison was accomplished with the hypergeometric test, which allows a statistical probability and Benjamini-Hochberg False Discovery Rate (FDR) correction[20]. This correction is essential to account for multiple testing. Encouragingly, many communities were significantly enriched (q-value ≤ 0.1) for a functional pathway (Reactome and Canonical Pathways), a disease pathway (DisGeNET), or a mixture of the two. Depending on the method, between 7-24% of communities were not enriched for any known pathway or disease and were regarded as novel. The exact breakdown of the community classification is shown in Figure 2A, and the majority of communities in BIGCLAM and RL are statistically overlapped with a function pathway and often with a disease pathway. Indeed, RL has the smallest fraction (7%) of novel communities, suggesting a higher positive predictive rate for the references. We noted that the number of genes in each community group generally increases from novel to mixed (Figure 2B). This could have a number of implications, including an observational annotation bias or a biological basis. These data show that community detection methods recover many commonly known functional and disease pathways but also discover new gene associations that possibly suggest novel pathways.

In order to assess the robustness of community detection we tested four metrics of

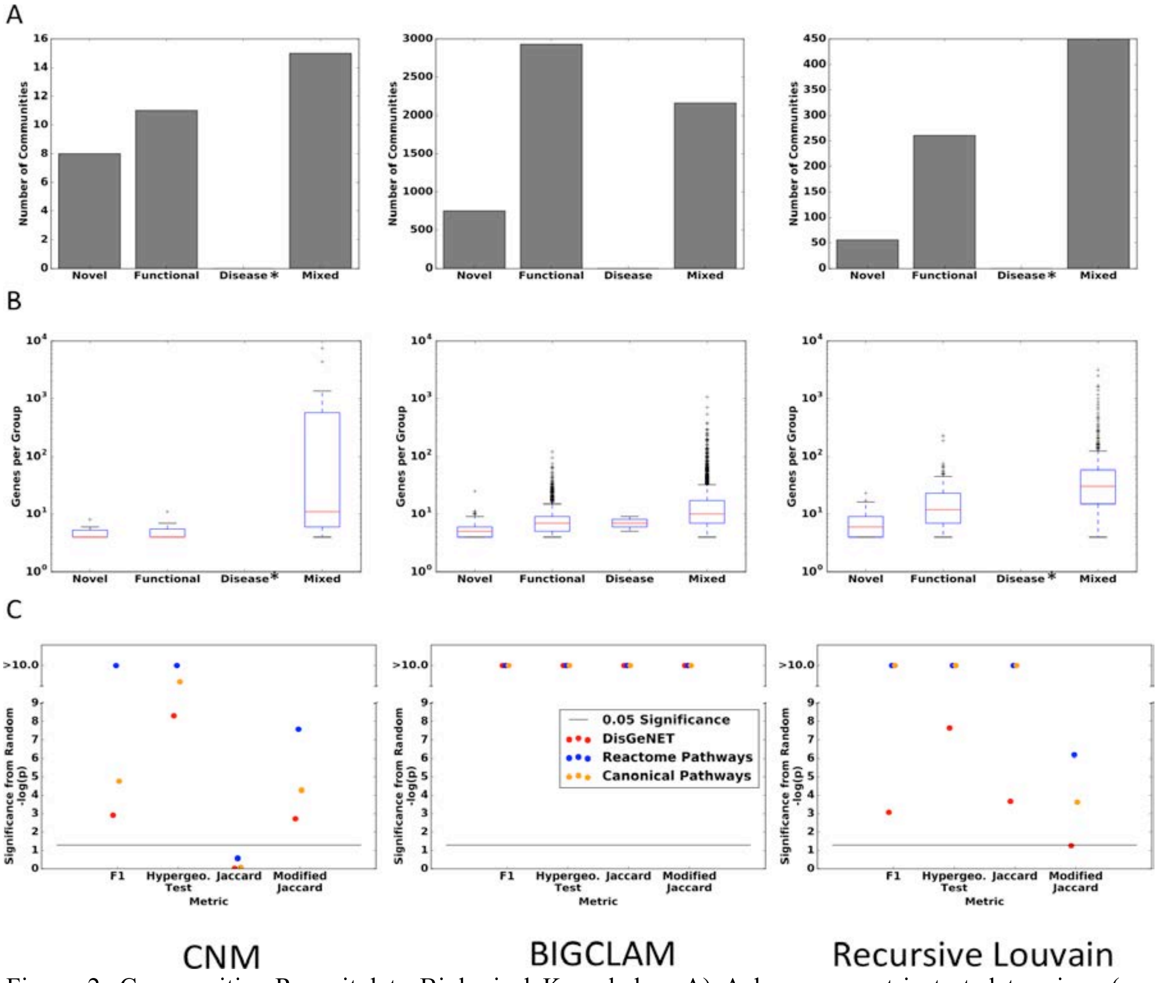


Figure 2: Communities Recapitulate Biological Knowledge. A) A hypergeometric test determines (q-value ≤ 0.1) whether a community is overlapped with a reference, and while many communities were overlapped with a disease or functional pathway, few or none (denoted by an *) were exclusively overlapped with only disease pathways. B) The number of genes in each group generally increases from a novel community to a community enriched for disease and functional pathways. C) All methods were non-randomly associated with every reference according to some metric, and many with p-values smaller than $10^{-10.}$

community overlap. Three random controls were generated to match the number and size of a set of communities and then scored against the references. Only the top score of a community or random against all pathways in a reference was kept. The distribution of random scores was then compared against the distribution of community scores using a Kolmogorov-Smirnov test. Due to poor performance and a lack of data (only 17 total communities), Louvain community detection (Supp. Figure 1-2) was assessed, but will not be shown because RL finds more total communities with overlap. As seen in Figure 2C, all three remaining methods were non-random by some metric; however, BIGCLAM and RL were significant on more metrics than CNM. In particular, BIGCLAM and RL appear highly significant in overlap with functional and disease pathways. RL has a higher percentage of communities that are enriched for both functional and disease pathways (Figure 2A), and this may suggest RL is better at recapitulating disease pathways. BIGCLAM has many more communities than the other methods (Figure 1A); this means we are more confident that BIGCLAM is performing different from random because we have more examples of overlap with the references. In contrast, Louvain community detection only found 17 communities, offering fewer opportunities to overlap with the references, and when we break those communities down further with RL, there is now more overlap with the references. These data show that BIGCLAM and RL recapitulate biological knowledge, while CNM appears to be less reliable.

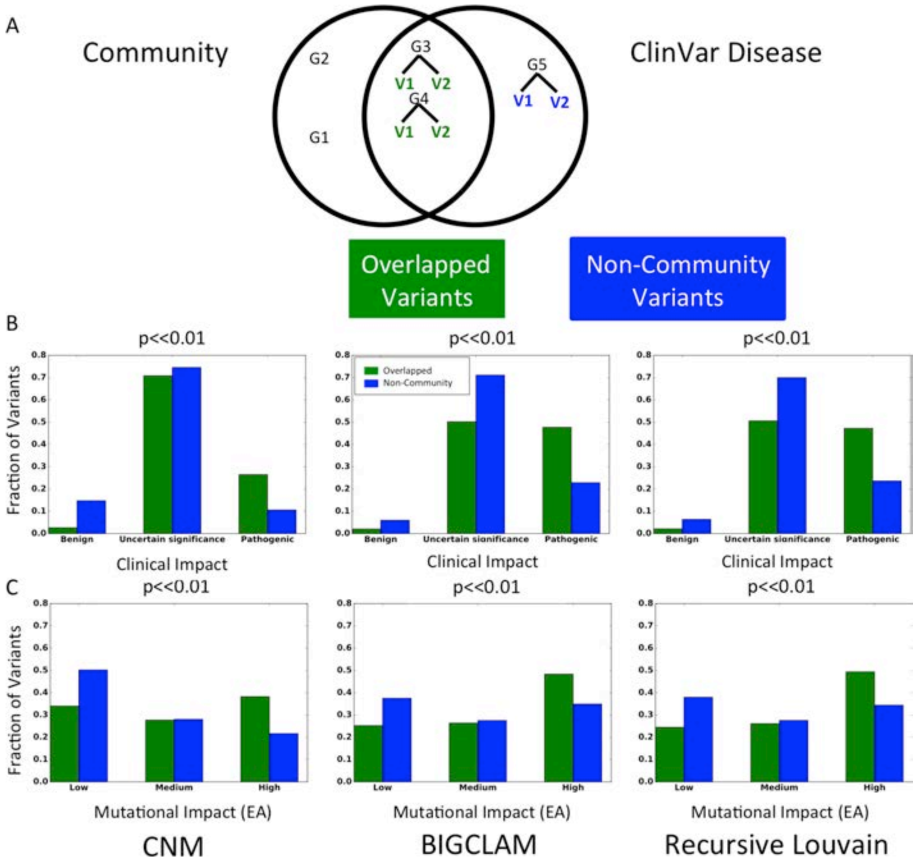


Figure 3: The Overlap between Communities and Diseases is Biased to Highly Pathogenic and Impactful Mutations. A) ClinVar groups variants into diseases. When a community and a disease from ClinVar both share genes, those genes possess a high B) clinical impact and C) mutation impact (p<<0.01) when compared to genes that are not found in communities. This implies that the communities are enriched towards variants that are pathogenic. Overlaps were only taken if the overlap was non-random (q<0.05).

2.3. *Clinical and disease relevance of communities*

A central question is whether these communities have real-world significance with respect to disease mechanisms. In order to address this question, communities were tested for overlap with diseases in the genetics database ClinVar. We hypothesized that communities represent units of biological function, and, if so, disrupting a gene that is part of a community would be more pathogenic than disrupting one that is outside of a community. Indeed, we find that mutations of disease genes that belong to communities have greater impact on the clinical phenotypes and on overall protein fitness

(Figure 3A). This was tested using the extensive annotations ClinVar[21] provides on the clinical impact of disease mutations. Specifically, for every community detection method, genes that fell inside communities showed are biased towards pathogenic variants (Chi-Square *p*-value << 0.01, Figure 3B). As an orthogonal control to test for bias in the impact of variations on disease genes, the Evolutionary Action (EA) provides an independent assessment of the deleterious impact of a mutation on protein fitness[22]. The same statistically significant trend emerges (Figure 3C, Chi-Square *p* << 0.01). These data show that mutations tend to have greater clinical and evolutionary deleterious impact if they affect genes that are part of communities.

To demonstrate a specific application of communities to disease pathways, we compared communities from BIGCLAM and RL, which outperformed CNM, against two diseases. These two diseases, Zellweger Syndrome (ZS) and Bardet-Biedl Syndrome (BBS) were both statistically associated with communities (*p* << 0.01). To associate diseases to communities, we used the disease-gene association information from two sources: (1) DisGeNet, a disease-gene association database integrating several public data resources and literature, and as shown in Figure 3, (2) ClinVar, a database providing the expert-asserted associations between genetic variants of genes and diseases. These disease-gene associations were used to calculate the statistical overlap between a disease and a community according to a hypergeometric distribution test of the overlap of genes, the unique genes of each, and all human genes. We hypothesized that when a community is statistically associated with a disease, any genes unique to the community are promising novel disease candidates. This hypothesis extends from a guilt-by-association assumption that has been successful in multiple systems[23,24]. As shown in Figure 4, when communities from multiple algorithms are compared to diseases, the overlaps possess high predictive power. For example, ZS is a peroxisomal biogenesis disorder characterized by severe hypotonia, epileptic seizures, and craniofacial abnormalities[25]. Because peroxisomal biogenesis depends highly on protein-protein interactions (PPIs), community detection on a PPI network reliably predicts and expands the disease definition. Indeed, using both community algorithms recovers all genes from DisGeNET and thirty additional genes are predicted. Of these thirty genes, fourteen are already annotated in the literature as being associated or causative of ZS. Two genes predicted to be associated with ZS are *ABCD1* and *ABCD2*, which are not known to be associated with ZS but transport very-long-chain fatty acids (*VLCFA*) across the peroxisome membrane

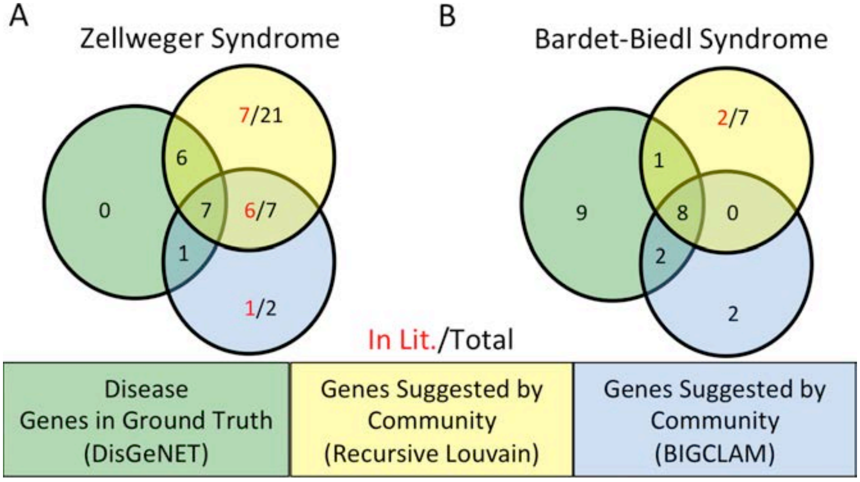

A
Zellweger Syndrome
7/21
6
0
7
6/7
1
1/2
B
Bardet-Biedl Syndrome
2/7
1
9
8
0
2
2
In Lit./Total
Disease Genes in Ground Truth (DisGeNET)
Genes Suggested by Community (Recursive Louvain)
Genes Suggested by Community (BIGCLAM)

Figure 4: Communities Discover Novel Disease-Gene Associations. A) For Zellweger Syndrome, all known disease associated genes in DisGeNET are recovered between a community from Recursive Louvain and BIGCLAM. Fourteen out of thirty predictions possess some form of evidence in the literature. B) Bardet-Biedl Syndrome (BBS) possesses significant overlap with the ground truth but does not find all known genes. However, there is a high concordance of overlap between the methods and the ground truth, with 2/7 of the Recursive Louvain predictions with literature evidence.

and cause adrenoleukodystrophy, a related peroxisomal disorder[26].

Another example is BBS, a rare ciliopathy that affects multiple body systems, where over half of the known genes were recovered and nine genes were predicted. Of these nine genes, two genes are already known in the literature to be associated with BBS. BBS is characterized by obesity, polydactyly, hypogonadism, intellectual disability, and renal abnormalities[27]. The gene *FOPNL* is suggested by community analysis but possesses no literature evidence. Despite this, *FOPNL* is well known to be associated with the biogenesis of cilia and BBS causative mutations upset cilial function. Furthermore, *FOPNL* interacts with *PCM1*, a known BBS gene that is also suggested by community analysis[28]. For BBS, there is a lack of overlap between community predictions, which points to the fact that each method is dependent on different features and therefore provides unique insight. These data demonstrate that communities can be useful in predicting and expanding sets of genes related to diseases that depend on protein interactions.

We determined if novel communities that lacked overlap with functional and disease pathways are biologically relevant by analyzing the co-expression of community genes in breast invasive carcinoma (BRCA). BRCA was chosen as a test case because it has a large number of patients with whom to power a co-expression study, though other cancers will be investigated in the future. If the genes in a community are co-expressed together more than randomly selected genes within tumor tissue RNA sequencing data, then that community represents a biologically relevant disease module. To validate our co-expression analysis, we examined four Reactome pathways, which are related to breast cancer pathways (PI3K/AKT activation, Signaling to RAS, PI3K/AKT Signaling in Cancer, and Constitutive Signaling by AKT1 E17K in Cancer) and found they are significantly co-expressed/regulated in breast diseased tissues (q<0.05). For both BIGCLAM and RL, at least 30% of the communities were co-expressed more than random with a q-value < 0.1 (FDR corrected by Benjamini-Hochberg), and over 52 % of novel RL communities were co-expressed non-randomly (Figure 5). Moreover, CNM preformed weaker than RL and BIGCLAM in comparisons to references, but with co-expression, CNM showed no signal, suggesting that it may be a poor approach for biological analysis. Overall, the figure shows that all classes of communities, including novel communities, have co-expression in BRCA.

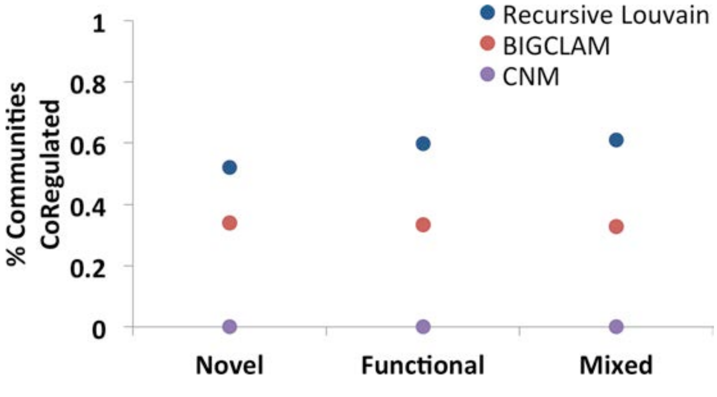


Figure 5: Communities are Significantly Perturbed in Cancer. In order to investigate whether novel communities had biological relevance, novel communities were investigated in the context of Breast Cancer (BRCA) co-expression data from TCGA. According to this analysis, the genes in novel communities are co-expressed to the same degree as communities with statistical overlap to functional and mixed pathways. This suggests that novel pathways represent a promising source of relevant biological knowledge.

As an application, one novel community (no. 657) detected by RL showed significant co-expression in BRCA (q = 0.00588) and has 14 gene members. Five members (*GPNTG*, *ECHS1*, *NACA*, *ABHD14B*, *NKX6-1*) were found to significantly coexist in the same subcellular location, extracellular vesicular exosome (GO:0070062; q = 0.01456; see Method). Furthermore, four members (*BTF3*, *GNPTG*, *CPEB2*, and *BICC1*) were found to be potentially co-regulated by the same transcription factor, *DACH1*, in a triple-negative breast cancer cell line, MDA-MB-231, (q = 0.0077 in ChIP-Seq enrichment analysis; see Method). Although it has been shown that *DACH1* expression level can predict BRCA survival[29] and play roles in breast cancer metastasis[30], *DACH1* currently has no pathway annotation in KEGG and Reactome databases. Therefore, this community might be the pathway related to *DACH1*. These results showed that novel communities could be related not only by expression but also by subcellular location and transcription factors. These data show the potential of communities to expand our knowledge of biology and disease.

3. Discussion

Determining the relationships between genes is essential for molecular biology and medicine. These relationships often cluster together into functional and disease pathways, and the characterization of these pathways is necessary to improve disease classification, patient stratification and, ideally, personalized treatment[5]. Here, we investigated the automated discovery of pathways by comparing several community detection algorithms against known functional and disease pathways and leading us to a novel application of the well-known Louvain algorithm, which we call Recursive Louvain (RL).

First, the communities detected by both BIGCLAM and RL were associated non-randomly with all the control, reference pathways. This strongly supports the biological relevance of these communities. Second, these communities also show a bias towards genes that experience pathogenic and high-impact variants in ClinVar. And third, regardless of the enrichment to a particular reference, these communities are often statistically co-expressed in breast cancer, including those that are new, in the sense that they are not enriched for any known functional or disease pathway. Therefore, these novel communities of genes may point to currently unrecognized biological pathways. Finally, in at least several cases, communities appear to predict genes associated with diseases with high predictive power. In the case of Zellweger Syndrome, six out of seven of the highest confidence predictions were already found in the literature although they were missing from the reference. The data from these approaches therefore consistently show that communities are biologically relevant.

The breadth of information in the input network limits community analysis. With only direct protein-protein interaction information, protein associations via indirect biological mechanisms such as transcription regulation can be missed. Eventually, the addition of transcriptional, post-translational, and epigenetic associations should help better characterize biological processes and extend the ability of community detection to recognize a wider variety of pathways. This is important as we note that, so far, many diseases and pathways are not enriched for communities. Beyond the breadth of information, community detection is also limited by its quality. Low-confidence, spurious associations between proteins surely lead to incorrect memberships of proteins in pathways. Furthermore, the pathways found represent global averages of associations. The future addition of context-specific transcriptional networks, such as from ChIP-seq data in ENCODE[31], should help find context-specific communities relevant to individual tissues or disease states. Despite these limitations, this work reveals the potential of topological network analysis in the identification and expansion of biologically meaningful pathways and shows that diverse results can be achieved through careful algorithm choice.

4. Methods:

4.1. ***Collection of reference sets****:* Reactome was downloaded from http://www.reactome.org, and was filtered for all disease pathways by trimming the disease section of the hierarchy as well as filtering out any pathway with the following words: disorder, hiv, defect, cancer, mutant, host, disease, influenza, toxin, viral, carcinoma, deletions, deficiency, variant, or virus. Canonical Pathways from the GSEA tool were downloaded from http://software.broadinstitute.org/gsea/downloads.jsp. Both KEGG and Reactome pathways are included in the Canonical pathways. All Reactome pathways in this dataset were filtered out to eliminate redundancy and then KEGG pathways related to diseases were filtered out to eliminate overlap with disease pathways from DisGeNET. DisGeNET was downloaded from http://www.disgenet.org/web/DisGeNET/menu/downloads as the curated dataset.

4.2. ***Community detection:*** Louvain community detection was calculated with the python community detection, which can be downloaded at http://perso.crans.org/aynaud/communities/. This base module then was used to create RL. RL runs Louvain, then takes each community larger than ten genes and makes it a subgraph of the original network, then calls Louvain community detection again. It does this iteratively until all communities have been broke down to ten genes or less or a gene has been seen in more than three communities. CNM and BIGCLAM communities were detected using implementations in the SNAP software package[32]. All community detection algorithms were applied onto STRING 9.1 experimental network[16].

4.3. ***Comparison to reference sets:*** All groups of genes were filtered to exclude pathways that contained three or fewer genes. This eliminated pathways that could easily be randomly recapitulated and therefore skew results. Pathways often overlap with each other, with minor differences between them. To prevent over counting from this, pathways that are too similar were collapsed together. Given a set of reference gene groups $R_i \in R$ and a set of community gene groups $C_i \in C$, all gene groupings were collapsed if the Jaccard Similarity > 0.9, where:

$$Jaccard\ Similarity, J(C_i, R_i) = \frac{|C_i \cap R_i|}{|C_i \cup R_i|} \tag{1}$$

To collapse two gene groups, the union of the genes was taken. In addition to the Jaccard Similarity, we then adopted three mathematical measures to evaluate the community detection algorithms outputs against the references, including a Modified Jaccard Similarity, a Hypergeometric Distribution test, and a F_1 score.

$$Modified\ Jaccard\ Similarity,\ J_m(C_i, R_i) = \frac{|C_i \cap R_i|}{|R_i|} \tag{2}$$

$$Hypergeometric\ Test, P(X \geq |C_i \cap R_i|) = 1 - \sum_{j=0}^{|C_i \cap R_i|-1} \frac{\binom{|R_i|}{j}\binom{M-|R_i|}{|C_i|-j}}{\binom{M}{|C_i|}} \tag{3}$$

$$F_1\ Score = \frac{1}{2}(F_R + F_C) \tag{4}$$

$$F_R or\ F_C = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \qquad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad , Recall = \frac{TP}{TP+FN} \qquad (6)$$

Where M=Number of genes in the original network and $F_R or\ F_C$ are the F_1 scores from the perspective of the reference or the community, respectively. TP represents the number of true positives; FP represents the number of false positives; FN represents the number of false negatives.

4.4 *Comparison to ClinVar:* In order to compare to ClinVar, we binned variants by clinical impact and Evolutionary Action, and the difference between each group of genes was assessed by a Chi-Square analysis. Only groups of genes from significant overlaps (q < 0.1 by hypergeometric analysis) between diseases and communities were assessed.

4.5. *Generation and evaluation of random controls:* Random controls were generated for each community set. For each community, a set of randomly generated genes were chosen from the protein interaction network such that the number of genes was identical to the number in the community. This was done three times in order to get a set of random communities that was then compared to the reference sets. The distribution of the random scores was compared against the distribution of the community scores using a Kolmogorov-Smirnov test. Each distribution was built with only the top score for a community or random against all pathways in a reference.

4.6. *Co-expression analysis in tumor tissues using RNA-seq data:* To determine if genes in a community have co-expression, RNA sequencing data version 2 of 1104 breast cancer tumor samples were downloaded from The Cancer Genome Atlas (TCGA) database (https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm) dated January, 2015. RNA-Seq by Expectation-Maximization (RSEM) normalized read counts (https://wiki.nci.nih.gov/display/TCGA/RNASeq) were used to represent mRNA expression level. The pairwise Spearman's rank correlation coefficients between the expression levels of pairs of genes in a community were computed. The distribution of absolute values of correlation coefficients was compared to the coefficient distribution of a random gene set, which is three times the size of a community, using a Kolmogorov-Smirnov test. All *p*-values were adjusted by Benjamini-Hochberg FDR correction[20]. A community was defined as co-expressed if the adjusted *p*-value is less than 0.1.

4.7. *Gene Ontology and ChIP-Seq enrichment analysis:* To understand the subcellular localization and potential upstream transcription factors of genes in a novel community, we analyzed the enrichment of Gene Ontology (GO) Cellular Component 2015 and ChIP Enrichment Analysis (ChEA) 2015 using Enrichr[33] (adjusted *p*-value < 0.1).

4.8. *Computation:* All calculations were done on an Ubuntu OS with 64 GB RAM and 4th Gen. Intel Core i7 3.7 GHz processor or equivalent machine.

4.9. *Supplemental data*: Supplemental data can be seen at:
http://mammoth.bcm.tmc.edu/SupplementalPSB2016Data.pdf

5. Acknowledgements

The authors would like to acknowledge the kind support of Christie Buchovecky, Daniel Konecki, Teng-Kui Hsu, and Panos Katsonis for their discussions and general feedback of the work. Additionally, the authors would like to acknowledge funding by NLM training fellowship (Grant No. T15 LM007093) for SJW, as well as funding from DARPA (N66001-14-1-4027), National Science Foundation (NSF DBI-1356569, NSF DBI-0851393), and National Institutes of Health (NIH-GM079656, NIH-GM066099).

References

1. Pawson, T. & Linding, R. Network medicine. *FEBS Lett* **582**, 1266-1270, doi:10.1016/j.febslet.2008.02.011 (2008).
2. AlQuraishi, M., Koytiger, G., Jenney, A., MacBeath, G. & Sorger, P. K. A multiscale statistical mechanical framework integrates biophysical and genomic data to assemble cancer networks. *Nat Genet* **46**, 1363-1371, doi:10.1038/ng.3138 (2014).
3. Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res* **44**, D481-487, doi:10.1093/nar/gkv1351 (2016).
4. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-462, doi:10.1093/nar/gkv1070 (2016).
5. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* **12**, 56-68, doi:10.1038/nrg2918 (2011).
6. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**, D447-452, doi:10.1093/nar/gku1003 (2015).
7. Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Mol Syst Biol* **3**, 88, doi:10.1038/msb4100129 (2007).
8. Clauset, A., Newman, M. E. & Moore, C. Finding community structure in very large networks. *Physical review E* **70**, 066111 (2004).
9. Yang, J. & Leskovec, J. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems* **42**, 181-213 (2015).
10. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008).
11. Ahn, Y. Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466**, 761-764, doi:10.1038/nature09182 (2010).
12. Palla, G., Derenyi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814-818, doi:10.1038/nature03607 (2005).
13. Taylor, I. W. *et al.* Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol* **27**, 199-204, doi:10.1038/nbt.1522 (2009).
14. Sah, P., Singh, L. O., Clauset, A. & Bansal, S. Exploring community structure in biological networks with random graphs. *BMC Bioinformatics* **15**, 220, doi:10.1186/1471-2105-15-220 (2014).
15. Yang, J. & Leskovec, J. in *Proceedings of the sixth ACM international conference on Web search and data mining.* 587-596 (ACM).

16. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* **41**, D808-815, doi:10.1093/nar/gks1094 (2013).
17. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res* **42**, D472-477, doi:10.1093/nar/gkt1102 (2014).
18. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).
19. Pinero, J. *et al.* DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database : the journal of biological databases and curation* **2015**, bav028, doi:10.1093/database/bav028 (2015).
20. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300 (1995).
21. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* **44**, D862-868, doi:10.1093/nar/gkv1222 (2016).
22. Katsonis, P. & Lichtarge, O. A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. *Genome research* **24**, 2050-2058, doi:10.1101/gr.176214.114 (2014).
23. Lee, I. *et al.* A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nat Genet* **40**, 181-188, doi:10.1038/ng.2007.70 (2008).
24. McGary, K. L., Lee, I. & Marcotte, E. M. Broad network-based predictability of Saccharomyces cerevisiae gene loss-of-function phenotypes. *Genome Biol* **8**, R258, doi:10.1186/gb-2007-8-12-r258 (2007).
25. Klouwer, F. C. *et al.* Zellweger spectrum disorders: clinical overview and management approach. *Orphanet J Rare Dis* **10**, 151, doi:10.1186/s13023-015-0368-9 (2015).
26. Burtman, E. & Regelmann, M. O. Endocrine Dysfunction in X-Linked Adrenoleukodystrophy. *Endocrinol Metab Clin North Am* **45**, 295-309, doi:10.1016/j.ecl.2016.01.003 (2016).
27. Khan, S. A. *et al.* Genetics of human Bardet-Biedl syndrome, an updates. *Clin Genet* **90**, 3-15, doi:10.1111/cge.12737 (2016).
28. Sedjai, F. *et al.* Control of ciliogenesis by FOR20, a novel centrosome and pericentriolar satellite protein. *J Cell Sci* **123**, 2391-2401, doi:10.1242/jcs.065045 (2010).
29. Wu, K. *et al.* DACH1 is a cell fate determination factor that inhibits cyclin D1 and breast tumor growth. *Mol Cell Biol* **26**, 7116-7129, doi:10.1128/MCB.00268-06 (2006).
30. Zhao, F. *et al.* DACH1 inhibits SNAI1-mediated epithelial-mesenchymal transition and represses breast carcinoma metastasis. *Oncogenesis* **4**, e143, doi:10.1038/oncsis.2015.3 (2015).
31. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100, doi:10.1038/nature11245 (2012).
32. Leskovec, J. & Sosič, R. SNAP: A General-Purpose Network Analysis and Graph-Mining Library. *ACM Transactions on Intelligent Systems and Technology (TIST)* **8**, 1 (2016).
33. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90-97, doi:10.1093/nar/gkw377 (2016).