Leveraging phenotype data to identify potential carriers of deleterious genetic variants

Theodore Morley B.A.¹, Lisa Bastarache M.S.², Nancy Cox Ph.D.¹, Douglas Ruderfer Ph.D.¹,²,³

¹Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA; ²Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA; ³Department of Psychiatry, Vanderbilt University Medical Center, Nashville, TN, USA;

Introduction

Heterogenous clinical symptoms caused by rare genetic variation may complicate identification of the underlying cause, prolonging diagnostic odyssey and delaying opportunity for potential medical benefits. Recent literature has shown that there exists a substantial population of individuals carrying deleterious genetic variants that go undetected in our current healthcare system, including those that cause known syndromes.[1] There is a clear need for us to maximize our ability to identify patients carrying genetic variants which could inform diagnosis, prognosis, and potentially treatment. We hypothesize that unique combinations of phenotypes seen infrequently in the population are likely to correspond to increased probability of an underlying genetic variant. Here, we aim to use only structured diagnostic information from electronic health records (EHR) in order to identify individuals at increased risk for carrying deleterious genetic events with the ultimate goal of validating these predictions through genetic assays.

Methods

We define our case population as 2,388 recipients of a form of genetic testing (chromosomal microarray) designed to identify large deletions and duplications of genetic material. These individuals were selected through analysis of patient notes within the Vanderbilt University Medical Center (VUMC) Synthetic Derivative. We then matched these cases in a one to four ratio with controls based on age, sex, race, record length within our EHR, and the number of unique years in which the patient had visited VUMC. We posed a prediction problem in which we aim to distinguish individuals who received genetic tests from those who did not, ideally replicating the suspicion which led clinicians to indicate which patients should receive genetic tests. We utilize a four-fold cross validation scheme on 80% of the data, and tested on the remaining 20%.

Utilizing only ICD9CM and ICD10CM codes translated to a vector of 1,685 unique pheWAS codes (phecodes) as input information, we considered two methods of representing these patients. The first of the methods consists of a binary matrix indicating presence or absence of phecodes, and the second is a broadly defined phenotypic risk score.[2] The phenotype risk score (pheRS) was originally defined as a disease-specific score indicating the likelihood of a given disorder based on the sum of the phenotypes specific to said disorder weighted by their log-inverse prevalence within the population. We instead calculated a pheRS for no specific disorder, creating a single score which aims to balance both the diversity of a patient’s phenotypes as well as the rarity of those phenotypes. For the binary phecode matrix we additionally evaluated three different methods of dimensionality reduction. They consisted of principal component analysis (PCA), uniform manifold approximation and projection (UMAP), and PCA fed into UMAP. Finally, we considered four different classification algorithms on this dataset; naïve Bayes, logistic regression, support vector machines, and random forest. After selecting a range of reasonable hyperparameters we applied a classic exhaustive grid search within our cross-validation framework and optimized our model selection on the F1 score, due to its balance of precision and recall.

For external validation with our best model, we identified a set of ~800,000 VUMC medical home patients, defined as having a record length of at least four years at VUMC. To assess the performance of our model, we identified 11,617 individuals within this sample who have at some point visited a genetic counseling clinic.

Results

Our sample included 2,388 case individuals who received genetic testing with chromosomal microarray, and 9,552 demographic matched controls. The case population is primarily very young, with 95% of individuals falling...
below 20 years of age. Additionally, the reported race information of the cases is as follows: 1728 white, 269 black, 204 unknown, 62 asian, 57 ‘other race’, 14 declined, and 5 ‘Alaskan/Indian’ (remaining individuals reported multiple races). We performed a pheWAS using chi-squared tests in order to identify the phenocodes which differed the most in prevalence between cases and controls. The most significant phenocodes found in this comparison represented largely developmental diagnoses such as delayed milestones, autism, and lack of normal physiological development which all had odds ratios > 10 and p-values < $10^{-200}$.

During four-fold cross-validation, the mean F1 score of all classifiers was found to be significantly higher when using the binary matrix of phenocodes as input, rather than the generalized pheRS. For example, the mean F1 score of logistic regression classification models using the binary matrix as input was ~0.63, while the same score for an input of the generalized pheRS was ~0.18. The F1 score was used as the optimization measure in our search, as opposed to PPV (precision). This was due to the fact that within the validation results, when optimizing on PPV, we found that the classifier with the highest PPV had extremely poor recall, compared to the classifier optimized for F1, while maintaining similar PPV. With respect to dimensionality reduction, the average classifier performance improved when any one of the three dimensionality reduction methods was applied. However, this average improvement came at the expense of the maximal performance. For our best performing classifier(random forest), the F1 scores of all classifiers given an input matrix reduced by PCA ranged from ~0.60 to ~0.67, while the same classifiers given the binary phecode matrix had F1 scores ranging from ~0.25 to ~0.71. This implies that optimal performance requires more complete information than was preserved during dimensionality reduction.

The best performing model selected through four-fold cross validation used a binary phecode matrix with no additional dimensionality reduction as input to a random forest classifier. Within the held-out test set this classifier had PPV of 0.73, NPV of 0.93, total accuracy of 0.89, and AUROC of 0.94. Further, our model is well-calibrated with a Brier score of 0.08. Projecting to an age comparable subset of VUMC medical home patients using genetic clinic visits as a target produced an AUROC of 0.86, suggesting that our model generalizes to the larger medical system.

**Conclusion**

As genetic testing becomes easier and clinical decision making based on genetic information becomes more valuable it is imperative that we are able to identify those patients most likely to benefit. Here, using a set of patients that received genetic testing and phenocodes we demonstrated the ability to predict these patients.

The performance of our model supports our initial hypothesis that unique combinations of diagnostic codes indicate the presence of a deleterious genetic variant. The success of an ensemble model such as random forest and the worsening of performance when the features are reduced or when using more linear approaches suggests complex patterns of phenocodes are present among these individuals. These unique constellations of phenotypes likely contribute to the challenge in identifying these patients clinically. Our current approach is most akin to building a model to replicate clinical suspicion since we depend on patients where providers previously believed genetic testing was warranted. This clinical suspicion will vary dramatically across clinical entities and likely has biases. Our ultimate aim is to refine our model to build an agnostic approach to optimally identify patients that carry clinically relevant genetic variants. Further work using more complete genetic data will be required for this effort.

We conclude that easily available phenotypic information within the EHR can provide evidence for those who should receive genetic testing. This evidence can potentially be used to both inform who should receive genetic testing so as to speed the diagnostic odyssey, as well as to prioritize selection of individuals for research purposes in a cost-effective manner.

**References**


McInnes, L, Healy, J. *UMAP*: Uniform manifold approximation and projection for dimension reduction. *ArXiv* e-prints 1802.03426, 2018