Name:	Mehr Kashyap, Martin Seneviratne
Affiliation:	Stanford University
Email:	mkashyap@stanford.edu, martsen@stanford.edu
Presentation type (s):	Lightning talk, poster

High-throughput phenotyping using imperfectly-labeled training data

Mehr Kashyap, BS[1*], Martin G. Seneviratne, MBBS[1*], Juan M. Banda, PhD[1], Nigam H. Shah, PhD[1]
[1]Stanford University, Stanford, California, United States

Abstract

The widespread adoption of electronic medical records provides a unique opportunity to accelerate research. To use these data for research, it is necessary to identify patients with phenotypes of interest and construct cohorts for observational studies. Traditional methods of phenotyping involve extensive rule-based definitions, which require a considerable amount of time to create. Recently, supervised learning techniques have been used to build phenotype definitions in the form of classifiers. The Automated PHenotype Routine for Observational Definition, Identification, Training and Evaluation (APHRODITE) is an R-package phenotyping framework that uses imperfectly labeled data to train classifiers, and has been shown to achieve good classification performance for type 2 diabetes mellitus and myocardial infarction. In previous work, the labeling heuristic used in APHRODITE have been based on mentions of phrases in textual data. In this study, we trained high-precision classifiers for four distinct phenotypes (type 2 diabetes mellitus, glaucoma, epileptic seizure, and peripheral vascular disease) using a labeling function based on multiple mentions of disease-specific codes. APHRODITE classifiers constructed with such imperfectly labeled data show significant improvements in recall compared to multiple-mention based labeling functions. These results demonstrate that it is possible to build phenotype classifiers in a high-throughput manner using the APHRODITE framework even in the absence of textual data to drive the labeling of the training set.

Introduction

Rapidly identifying patients within an electronic medical record (EMR) who exhibit a particular phenotype, such as having type 2 diabetes mellitus (T2DM), is necessary for cohort-building in observational studies [1]. Due to the heterogeneity of EMR data, this phenotyping task presents a significant challenge. Traditional methods for electronic phenotyping involve building detailed database queries manually constructed by clinicians, which are time-consuming to create [2]. Recent work has shown that supervised learning approaches can be used to build classifiers for electronic phenotyping in a rapid and generalizable format [3][4].

In the Observational Medical Outcomes Partnership (OMOP) collaborative group, the Automated PHenotype Routine for Observational Definition, Identification, Training and Evaluation (APHRODITE) R-package has been used to build phenotype classifiers using imperfectly-labeled data. In previous work, both Agarwal *et al.* and Banda *et al*. created their training sets based on the presence of a set of words or a phrase in clinical notes, to create classifiers for both T2DM and myocardial infarction [5][6].

This paper explores an alternative imperfect labeling function based on multiple mentions of disease-specific codes. The hypothesis is that in the absence of textual data, using multiple mentions of disease codes will yield a training set with high precision at the expense of recall; and that during classifier training, the learning algorithm will be able to generalize such that the final model will have higher recall than that of the original labeling function. We test this hypothesis by constructing phenotyping classifiers for four conditions with the goal of improving recall while maintaining precision.

Methods

We selected four phenotypes (type 2 diabetes mellitus, glaucoma, epileptic seizure, and peripheral vascular disease) for which rule-based definitions were created by the OHDSI community at the hackathon on 3/17/2017. For each phenotype, we identify cases by searching patients' EMR data for at least five mentions of a disease specific SNOMED code associated with the phenotype of interest. We compared the performance of this heuristic "as-is" with a model built using APHRODITE, an R-package that uses the imperfectly-labeled data to train classifiers. Specifically, we trained L1 penalized logistic regression models using 5-fold cross validation on patients labeled with the 5-mentions search.

The rule-based definitions created at the hackathon identify the set of patients comprising the evaluation set. Since these definitions were developed and validated by several members of the OHDSI community, they serve as a good surrogate for manually reviewing patient data to create evaluation sets. Each test set was composed of 5000 patients, with the proportion of cases set equal to the population prevalence of the corresponding phenotype.

Results

Requiring cases to have five mentions yielded perfect precision, as it is improbable that any patient with five mentions of a code does not have the associated phenotype. Achieving high precision, however, results in drastically low recall. The mean recall for requiring at least five mentions was 0.06 (Table 1).

Phenotype	Prevalence used in training set	5 mentions of SNOMED code		APHRODITE classifier		Recall boost using a classifier	Precision loss using a classifier
		Recall	Precision	Recall	Precision		
T2DM	0.14	0.11	1.00	0.39	0.92	**0.28**	**0.08**
Glaucoma	0.01	0.04	1.00	0.24	1.00	**0.20**	**0.00**
Epileptic Seizure	0.017	0.02	1.00	0.20	0.85	**0.18**	**0.15**
PVD	0.052	0.08	1.00	0.20	0.95	**0.12**	**0.05**

Table 1. Test set performance of 5-mentions code search, and APHRODITE classifier trained on data labeled with the 5-mentions search (n = 5000 with cases:controls ratio set at population prevalence)

Logistic regression models using data labeled with the 5-mentions search improved recall, while only marginally reducing precision. Specifically, classifiers showed a mean boost in recall of 0.20 with a mean loss in precision of 0.07 compared to the 5-mentions search. These results indicate that it is possible to create phenotype classifiers that have better recall than that of the labeling function which uses multiple code mentions.

Conclusion

This study demonstrates the use of multiple mentions of disease-specific codes to create labeled training data, and the ability to train logistic regression classifiers on those data using APHRODITE. Across four phenotypes, APHRODITE classifiers show considerable gains in recall relative to using multiple mentions of a code. These classifiers maintain high precision despite the low population prevalence of each disease. The classifiers are also significantly faster to generate (requiring under two hours of computational time) than manually created rule-based phenotype definitions, and are readily transferrable to other EMR datasets mapped to the OMOP CDM. This high-throughput phenotyping methodology enables creation of phenotype definitions "on-demand" for identifying patients eligible for a research study across a large, heterogenous EMR dataset.

References

[1] C. A. Longhurst, R. A. Harrington and N. H. Shah, "A 'green button' for using aggregate patient data at the point of care," *Health Aff (Millwood),* vol. 33, no. 7, 2014.

[2] K. M. Newton, P. L. Peissig, K. A. N, S. J. Bielinski, R. L. Berg, V. Choudhary and et al., "Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. Journal of the American Medical Informatics Association," *JAMIA,* vol. 20, no. e1, 2013.

[3] S. Yu, K. P. Liao, S. Y. Shaw, V. S. Gainer, S. E. Churchill, P. Szolovits, S. N. Murphy, I. S. Kohane and T. Cai, "Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources," *JAMIA,* vol. 22, no. 5, 2015.

[4] Y. Halpern, S. Horng, Y. Choi, D. Sontag, "Electronic medical record phenotyping using the anchor and learn framework," *JAMIA,* vol. 23, no. 4, 2016.

[5] J. M. Banda, Y. Halpern, D. Sontag, N. H. Shah, "Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network," AMIA Jt Summits Transl Sci Proc.*,* p48-57

[6] V. Agarwal, T. Podchiyska, J. M. Banda, V. Goel, T. I. Leung, E. Minty, T. E. Sweeney, E. Gyan and N. H. Shah, "Learning statistical models of phenotypes using noisy labeled training data," *JAMIA,* vol. 23, pp. 1166-73 , 2016.