Lift your Anchors and Begin the OHDSI with APHRODITE

Yoni Halpern[1], Steven Horng, MD, MMSc[2], David Sontag, PhD[2], Juan M. Banda, PhD[3], Nigam Shah, MBBS, PhD[3]
[1]New York University, New York, NY;
[2]Beth Israel Deaconess Medical Center, Boston, MA;
[3]Stanford University, Stanford, CA.

Abstract

We describe a novel method of defining electronic medical record phenotypes using classifiers learned from large collections of imperfectly labeled patient records. The method allows for building scalable libraries of phenotype definitions that can be used for a variety of purposes, from real-time alerts, to population targeting, to cohort selection for observational studies.

Introduction

Extracting phenotype information from electronic medical records is extremely useful for retrospective observational studies (1) (2), as well as generating real-time alerts and reminders in the clinical setting. Traditional methods of identifying EMR phenotypes involve the consensus construction of inclusion and exclusion criteria (3). Recent work uses statistically learned classifiers to extract positive cases, but these generally rely on a gold-standard labeled set. Therefore, expanding the library of phenotypes requires an intensive commitment of time from domain experts.

Learning with Noisy Labels

In our work, we suggest a different route, which we call learning with anchor variables(4) or silver-standard labels (5). We take a statistical machine learning approach, where the phenotype definition is described by a classifier learned from a large collection of historical medical records. However, rather than require that a domain expert labels cases and controls, the expert specifies a simple set of conditions that are sufficient to identify a subset of the positive cases in the data (e.g. rules with high positive predictive value but not necessarily high sensitivity).

These conditions can use structured data from the EMR (e.g., medications such as metformin as an indication of diabetes, lab results like EFGR for chronic kidney disease or a positive rapid strep antigen test for strep throat), or unstructured textual data (e.g., "From nursing home" to indicate whether the patient is from a nursing home, or the presence of textual terms already associated with the concept in the UMLS).

These rules are then used to create an imperfectly labeled dataset that is suitable for applying standard machine learning classification techniques such as logistic regression or random forests. This method allows us to learn from large unlabeled datasets, and the large number of samples more than compensates for the noise in the labeling procedure. (6)

Aphrodite

Choosing cohorts using an anchoring strategy has been implemented within the OHDSI framework as the Automated PHenotype Routine for Observational Definition Identification Training and Evaluation (APHRODITE) package. We have constructed and evaluated phenotype models for four conditions using APHRODITE (Tables 1,2) and the tool is released on the OHDSI github site at: https://github.com/OHDSI/Aphrodite.

Abdominal pain	Cerebrovascular accident	Kidney stone
Alcoholism	Diabetes	Laceration
Allergic reaction	Deep vein thrombosis	Liver (history)
Ankle fracture	Employee exposure	Motor Vehicle Accident
Anticoagulated	Epistaxis	From nursing home
asthma/copd	Gastroenteritis	Pancreatitis
Back pain	Gastrointestinal bleed	Pneumonia
Bicycle accident	Geriatric fall	Psych
Cancer	Headache	Obstruction
Cardiac etiology	Hematuria	Septic shock
Cellulitis	Hiv+	Severe sepsis
Chest pain	Intracerebral hemorrhage	Sexual assault
Congestive heart failure	Immunosuppressed	Suicidal ideation
Cholecystitis	Infection	Syncope
		Urinary tract infection

Figure 1: (Left) Phenotypes currently being identified in real-time at BIDMC. (Right) Display screen where one of the phenotypes has been used to recommend a pathway of care.

	Aphrodite		Consensus Rules	
Phenotype	**Accuracy**	**PPV**	**Accuracy**	**PPV**
Type II Diabetes	0.93	0.83	0.92	0.96
Myocardial Infarction	0.92	0.92	0.87	0.84

Table 1 Performance of phenotype models learned with Aphrodite compared to standard consensus rules for Type II Diabetes (7) and Myocardial Infarction (OMOP broad definition with hospitalization available at omop.org/hoi) for extracting phenotypes from Electronic Medical Records.

Phenotype	AUC	Sens.	Spec.	PPV
Familial Hyperlipidemia	0.90	0.765	0.936	0.2
Celiac disease	0.75	0.40	0.90	0.04

Table 2 Performance of two new phenotype models learned with Aphrodite that do not have comparable rule-based definitions.

Interactive selection of Anchors

Specifying anchors can be difficult. We implement a feedback mechanism that allows users to refine their choice of anchors over time. A logistic regression model with high L1 regularization is learned on the imperfectly labeled dataset, and the non-zero weights are presented as suggestions for possible additional anchors. Suggested anchors are accompanied by a list of patients who were not initially considered positive cases, so that the impact of adding each new anchor can be assessed. The user then evaluates whether the new anchor would be a useful addition and has the option of adding the anchor and retraining the classifier. In this way, we allow starting with a small initial set of anchors, and iteratively refining the anchors to create a more comprehensive set of positively labeled cases to be used to train the final classifier. The use of additional anchors can reduce the error rate of the imperfect labeling—both in terms of increasing precision and recall.

Learned Phenotypes and Use Cases

The learning with anchors pipeline has been implemented both at Stanford Medical School and at Beth Israel Deaconess Medical Center (BIDMC). At BIDMC, 43 different phenotype definitions are being used in real-time to trigger alerts in the emergency department and suggest patient enrollment in standardized "pathways of care" (8) (Figure 1). At Stanford, phenotype definitions for Type II Diabetes and Myocardial Infarction were compared to manually created rules and found to have comparable sensitivity and specificity, while requiring a fraction of the time to build. In addition, a phenotype definition for Familial Hyperlipidemia is being used in the FIND (Flag, Identify, Network, Deliver) FH project to identify individuals with the disorder who are undiagnosed, untreated or undertreated. Comparing between the Type II Diabetes model at Stanford and the diabetes model at BIDMC (which includes both type I and II diabetes) revealed that many similar intuitive features appeared in both classifiers. For example, the highly weighted features in both models contained textual mentions of diabetes, high glucose readings, and medications such as metformin.

Open Question: Cohort Visualization

Interactively building a cohort definition requires a mechanism to summarize a patient cohort as well as review indicial records to evaluate whether the cohort is reasonable and to determine whether more/different anchors are needed. When textual elements are included in the patient representation, it is possible to scan through patients and quickly evaluate the quality of the patients selected into the cohort. In the OHDSI framework where text elements are relatively rare, it is a significant challenge to evaluate the utility of suggested anchors, and the impact of including them to find cases for model building.

Conclusion

Learning with a large number of noisy labels presents an alternative, effective method of quickly building phenotype definitions that perform competitively with manual collection of a smaller number of "gold standard" labeled data, and requires a fraction of the time. Such rapid phenotyping approaches are essential for the success of large scale collaborative projects such as OHDSI that will need to build phenotype libraries with hundreds of phenotypes.

References

1. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. Journal of the American Medical Informatics Association : JAMIA. 2013;20(e2):e226-e31.
2. Denny JC, Ritchie MD, Crawford DC, Schildcrout JS, Ramirez AH, Pulley JM, et al. Identification of Genomic Predictors of Atrioventricular Conduction: Using Electronic Medical Records as a Tool for Genome Science. Circulation. 2010;122(20):2016-21.
3. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc. 2013;20(e1):e147-54.
4. Halpern Y, Choi Y, Horng S, Sontag D, editors. Using Anchors to Estimate Clinical State without Labeled Data. American Medical Informatics Association (AMIA) Annual Symposium; 2014.
5. Agarwal V, Lependu P, Podchiyska T, Barber R, Boland M, Hripcsak G, et al., editors. Using narratives as a source to automatically learn phenotype models Workshop on Data Mining for Medical Informatics; 2014.
6. Natarajan N, Dhillon I, Ravikumar P, Tewari A. Learning with noisy labels. Advances in Neural Information Processing Systems 26. 2013.

7. Wei WQ, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. J Am Med Inform Assoc. 2012;19(2):219-24.
8. Rotter T, Kinsman L, James E, Machotta A, Gothe H, Willis J, et al. Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs. The Cochrane database of systematic reviews. 2010(3):Cd006632.