FIND FH – A phenotype model to identify patients with familial hypercholesterolemia

Katherine E. Niehaus, MS[1], Juan M. Banda, PhD[2], Joshua W. Knowles, MD, PhD[2], and Nigam H. Shah, MBBS, PhD[2]
[1]University of Oxford, Oxford, UK; [2]Stanford University, Stanford, CA, USA

Abstract

Familial hypercholesterolemia (FH) is a treatable but frequently overlooked genetic condition that leads to abnormally high levels of low density lipoprotein (LDL) cholesterol in the blood. In previous work, we have shown the feasibility of using noisy labeling to create "silver standard" training sets to build statistical phenotyping models. We describe an Electronic Health Records based phenotyping model for identifying patients that may have FH. We discuss the potential use of our model for identifying FH patients in a cardiovascular clinical setting, and highlight the importance of choosing the control population of patients accordingly. The goal of our efforts is to identify FH patients for the purpose of alerting physicians and improving care, rather than identifying a cohort of patients upon which to perform further analysis. The potential of our algorithm for benefitting patients is high. With a well-performing model and a functional outreach program, such as Flag, Identify, Network, Deliver (FIND) FH, we can contact physicians whose patients may have undiagnosed FH.

Introduction

Familial hypercholesterolemia

Familial hypercholesterolemia (FH) is a genetic condition that causes alterations in cholesterol metabolism, resulting in abnormally high levels of low density lipoprotein (LDL) cholesterol in the body (1). Because LDL causes atherosclerosis, and given the extremely high exposure throughout life in untreated FH individuals, these patients have a risk of myocardial infarction twenty times that of the general population (2). However, patients' cholesterol levels can return to normal with aggressive management. Historically, statins have been the primary treatment, although the recent approval of PCSK9 inhibitors now provides an alternative as well.

While treatable, FH is frequently overlooked; in a 2011 survey of 500 cardiologists, fewer than 30% were able to recognize a case example (2). Furthermore, there is no ICD9 code for FH, making it largely invisible to the healthcare system. Indeed, while repeated studies have found the prevalence of FH to be between 1 in 200 and 1 in 500, the current diagnosis levels suggest that less than 1% of patients with the condition in the US have been diagnosed (1). Given the significant increase in cardiovascular risk, together with the treatable nature of the condition, it would be very desirable to identify individuals who likely have FH, but who are currently undiagnosed. One way to accomplish this is to identify patients whose electronic health record (EHR) "signature" is similar to patients known to have FH – i.e., to create a mechanism to identify the latent patterns of term mentions, lab test orders, medication usage that is indicative of undiagnosed FH patients. The FH Foundation's Flag, Identify, Network, Deliver (FIND) FH initiative, launched in 2015, is an partnership between Amgen, the American Heart Association, Stanford University, and the FH Foundation intended for this purpose.

Phenotyping

The identification of patients with a specific condition from the EHR is a challenging problem because EHR data is often incomplete, has misreported values, and is coded differently across hospitals. As a result, there have been many research efforts to create systematic procedures (often called "algorithms") to identify such patient groups with known medical conditions – we refer to such efforts as electronic phenotyping. Approaches range from using combinations of query terms to supervised and semi-supervised machine learning approaches (3–6). The eMERGE network, for instance, has developed a number of specific rule-based algorithms for various disease conditions, which are maintained on the Phenotype KnowledgeBase (PheKB) (7,8). These rule-based approaches are usually designed by committees of experts. Increasingly machine learning approaches are being employed, which allow for automated learning of an algorithm but still rely upon labeled training examples. In both cases, these approaches represent an attempt to automatically identify patients with a specific known condition and are different from exploratory unsupervised approaches, which aim to identify novel phenotypes and sub-phenotypes (9–11).

In previous work, we have shown the feasibility of using noisy labeling to create "silver standard" training sets to build phenotyping models for several medical conditions (12,13). This approach represents a nearly automated

phenotyping methodology because cases and controls do not have to be manually labeled. Instead, a highly specific concept (e.g. "type 2 diabetes mellitus" for Type 2 diabetes) is used to generate synonym concepts, and patients with these concepts mentioned in their record are defined as cases. Controls are patients who lack these concepts. This approach can be viewed as an example of anchor based learning proposed by Halpern et al. (14). However, some conditions, such as FH, are very difficult to extract using the silver standard approach because they are poorly codified in the record. The term "familial hypercholesterolemia" is rarely written in the notes, and as previously mentioned, no specific ICD9 code exists. Development of a classification model in this situation, therefore, does require manual identification of some gold-standard true-positive cases.

An additional challenge in building an FH phenotyping model is the very small number of labeled positive cases that a single institution can provide (due to the low diagnosis rate). In addition, as is the case for any phenotyping model, the way in which the different data types found in the EHR are combined into features must be carefully considered. Here, we present our initial model for identifying patients that may have FH. We perform feature selection within each feature class individually, and combine the resulting features into a single model. We discuss the potential use of our model for identifying FH patients in a cardiovascular clinical setting, and highlight the importance of choosing the control population of patients accordingly.

Methods

Data

To train our model, we used data from the Stanford Translation Research Integrated Database Environment (STRIDE), which consists of data collected between 1994-2013 from Stanford Hospital and Clinics and the Lucille Packard Children's Hospital. All data was formatted into the Observational Health Data Sciences and Informatics (OHDSI) Common Data Model (CDM) framework, which is an international effort to create a standardized format for observational healthcare data. Our data consisted of four main feature types: patient lab measurements; drug prescriptions; ICD9/CPT codes; and terms from doctors' notes.

Patient population

In collaboration with a cardiologist who specializes in FH, we identified 71 FH patients in our dataset. These patients composed our gold-standard case group. The choice of a control group of patients depends upon the intended use of the model. Because the proposed initial setting is within cardiovascular clinics and because the most difficult cases to distinguish are FH versus high-cholesterol (but non-FH) patients, we use high cholesterol (HC) patients as our control group. Assuming a prevalence of 13.1% of high cholesterol in the general population (15), coupled with a 1/350 prevalence of FH (1), we estimated the ratio of FH:HC in a typical cardiovascular clinic to be about 1:46. To create a pool of HC patients, we used a noisy-labeling approach by identifying all patients with the term "hypercholesterolemia" in their notes (n=35,000). Because we wanted to assess the performance of our model in real-life prevalence conditions, our test sets (within cross-validation folds, as described below) consisted of 642 HC patients and 14 FH patients. To minimize the degree of unbalance while training our classifier, we drew 300 HC patients and included the remaining 57 FH patients for training (within cross-validation folds). We did not conduct a manual chart-review of the HC patients to ensure that they did not have FH. Assuming an FH prevalence of 1/38 within HC patients, and a worse-case scenario that none of these FH patients have been identified, at a maximum we would expect that 6-7 patients from the set of 300 HC patients could potentially have FH.

Model Construction

We approached our model development in a data-driven way, as we did not a priori know which aspects of the patient record may be discriminative for FH. Therefore, to capture all possible informative aspects of the record, our features consisted of normalized counts of all unique concept mentions in a patient's EHR. By concepts, here we mean any unique code, drug, or term from the notes. For laboratory measurements, we split each lab test into a categorical encoding, where we defined a lab concept as a particular lab type, plus the label of being low, normal, or high (where low, normal, and high cutoff points were determined by the clinical reference ranges for the given test). For each of the four feature types, therefore, we obtained raw counts of the number of times (counting only once per visit) a given code, drug, term, or lab level appeared across a patient's record.

However, some patients may have records that cover several years, while others may only have a single visit. To account for these differences, we normalized each patient's counts by the total number of features they had within that feature type. For instance, if a patient's record has ten occasions when a high cholesterol level is recorded, and forty unique labs ever recorded across the record, the feature value would be 10/40 = 0.25. This method of normalization essentially provides emphasis to those terms/labs/codes/drugs that dominate the patient's record. The intuition for such normalization is that cholesterol-related concepts will be the primary repeated characteristic in the record for patients with FH.

The content of information contained in each of our feature types varies considerably – the notes and codes data in particular are very sparse, as most patients only have instances of a small subset of possible codes or terms. In addition, one of the goals of our model is that it be easily transferrable to other EHR systems, meaning that a reduction in the number of features required would be desirable. Taking these facts into account, we devised a feature selection schema.

Our feature selection process was placed within a nested 5-fold cross-validation loop so that the variance in classification performance could be assessed across different slices. The steps performed within each cross validation fold are outlined in Figure 1. Within each cross-validation fold, the case patients (N=71) were split into 80%:20% training:testing. From the control patients, 642 were set aside as a test set, to match the estimated prevalence of FH patients amongst high cholesterol patients. Then, for each feature type, we made ten draws of M=300 control patients. We selected just the features of that type for the 57 training cases and the 300 controls in each of these draws, and used these patients to construct a random forest (RF) model to classify FH patients. We chose the RF model because of its easily-interpretable feature weightings (in contrast, for example, to radial basis function support vector machines). We therefore created ten models for each feature type. Using these models, we identified a limited feature set of consistently discriminative features: those features found in the top 50 features in at least nine of the ten models.

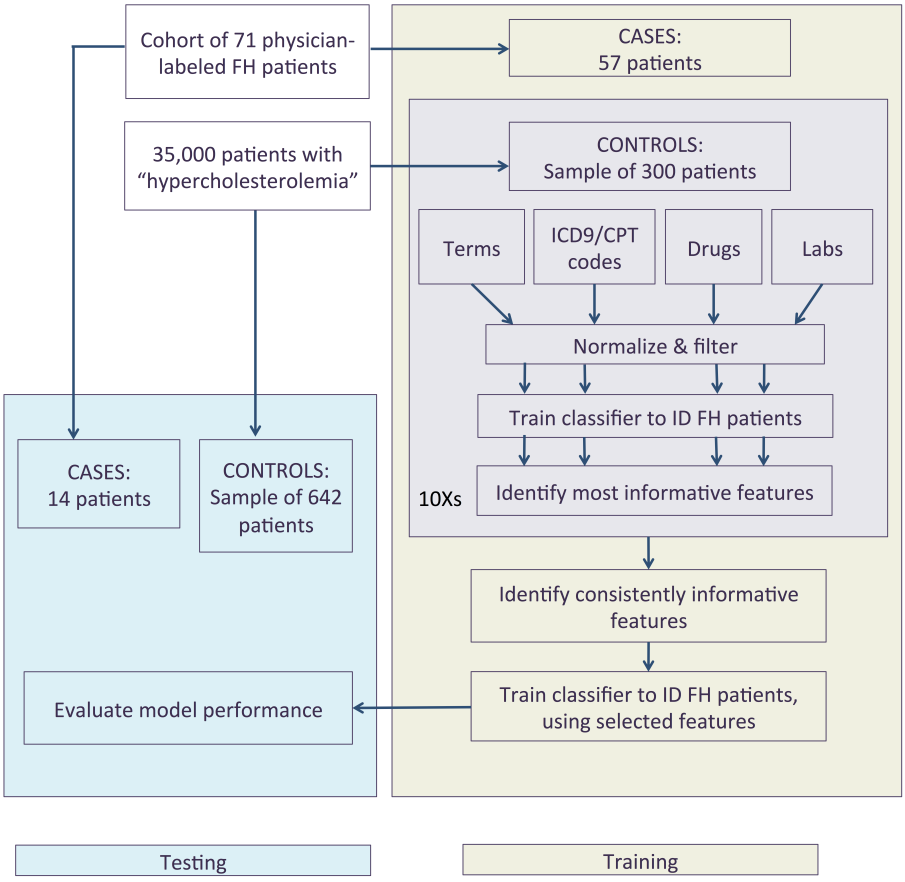

Figure 1. Description of feature selection process that occurs within model training during each fold of nested cross-validation.

We combined the limited feature sets from the four feature types to create a final feature set, which we used to train a final RF model using another draw of 300 control patients and the 57 case patients. In training all of our RFs, rather than optimizing for accuracy, we optimized the F-score, with $\beta=5$ (to reflect the approximately 5-fold class imbalance in training). The F-score is metric that combines the precision and recall achieved by the classifier, where the β parameter determines the degree to which one class's performance is weighted over another. We also performed internal five-fold cross-validation to optimize the model parameters (for RFs, this was the number of features for each tree). Across all feature types, before any feature selection was performed, we eliminated any features found in less than 10% of patients. Again, all of these steps are performed within the training set of just a particular fold.

We will refer to the model we have described above as our "default" model; as a sensitivity analysis, we also experimented with removing various feature classes from the model; employing no feature selection at all; using regularized logistic regression rather than RFs; and comparing different normalization techniques. Given the class imbalance in our testing set, it was inappropriate to use a measure such as accuracy to assess the model's classification performance. We instead used the area under the ROC and precision-recall curve to evaluate our models. We used Aphrodite (16), an R package we have developed for phenotyping with silver standards, as the basis for much of this work.

Results

Combining results across our five cross-validation folds, we found that we were able to classify our held-out cohorts of 14 FH and 642 HC patients with a mean AUROC of 0.90 (standard deviation (sd)=0.02). The area under the precision/recall curve for correctly identifying FH patients was 0.294 (sd=0.091), and the area under the precision/recall curve for correctly identifying non-FH patients was 0.997 (sd=0.00051) (see Figure 2). This means that, pooling across all held-out folds and taking a probability threshold cutoff of 0.5, 43 of the total 71 FH patients would have been correctly identified by the model. Across all runs, 138 of the >3,000 HC patients would have been incorrectly classified as having FH, and 28 true FH patients would have been missed. In contrast, in the model without feature selection, at the same probability cutoff, only 36 FH patients would be correctly identified, but at the same time, only 90 HC patients would be incorrectly classified. In the default model, this can be restated as a 61% sensitivity being associated with a 24% positive predictive value (as shown in the precision-recall curve).

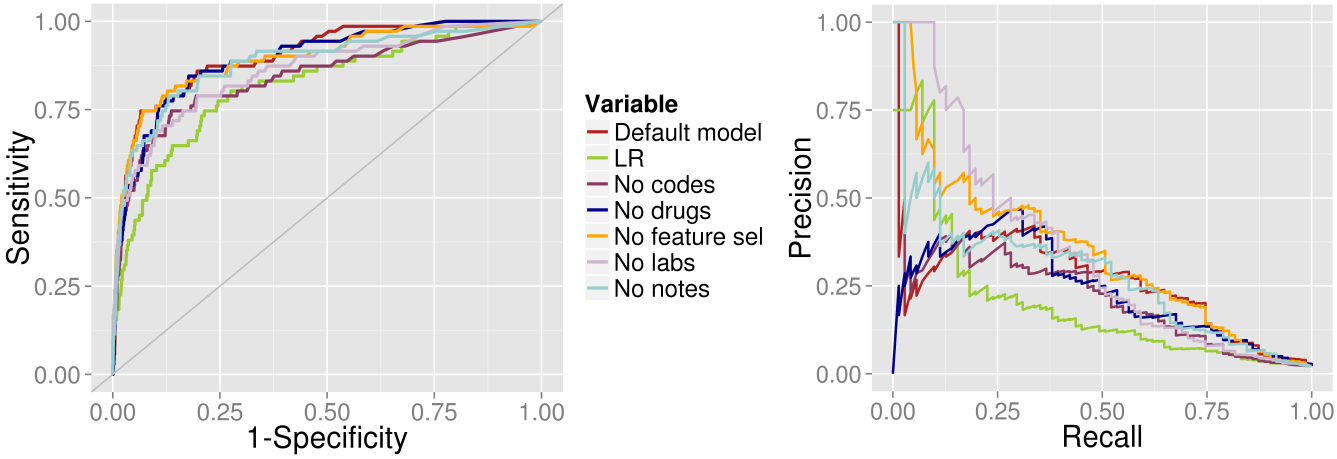

Figure 2. ROC and PR curves for the various models considered

We found that classification performance dropped most precipitously when we removed the coded class of features from the model, while the performance was least affected by the removal of drug features. Table 1 provides the full details of the results of various modifications we have made to the model. These findings make sense when examined in light of the most discriminative features, as will be presented below.

Table 1. Performance metrics for various models that were examined. Largest values are in bold.

	Area under ROC	Area under PR curve	Sensitivity at 0.5 cut-off	Specificity at 0.5 cut-off
Default model (feature selection + RF)	**0.905 (0.015)**	0.294 (0.091)	60.6%	95.7%
Logistic regression model	0.822 (0.066)	0.227 (0.15)	56.3%	90.9%
No feature selection	0.896 (0.041)	**0.366 (0.34)**	50.7%	**97.1%**
Note features removed	0.887 (0.061)	0.305 (0.050)	**63.3%**	95.5%
Code features removed	0.846 (0.061)	0.244 (0.13)	52.1%	95.5%
Lab features removed	0.864 (0.032)	0.358 (0.061)	49.3%	**97.1%**
Drug features removed	0.895 (0.038)	0.311 (0.14)	53.5%	95.5%
Just notes and codes	0.867 (0.069)	0.328 (0.065)	62.0%	94.8%
No normalization	0.883 (0.046)	0.349 (0.16)	50.7%	96.2%
Normalization by months of follow-up	0.876 (0.048)	0.353 (0.077)	59.2%	95.1%
Normalization by number of visits	0.895 (0.051)	0.295 (0.080)	56.3%	96.0%

Selected features

The total number of features included in the model prior to feature selection was 1187. Across the five folds of the default RF model, the number of features that were kept following the feature selection step ranged from 152 to 170. The features that were found to be most discriminative in the final models (i.e., after feature selection) for FH are shown in Table 2; the class in which these concepts were enriched is shown. Additional features outside these top ten included atorvastatin (drug feature), lipids and triglycerides (note features), coronary arteriosclerosis (code), and age. As is evident, physicians appear to be using the ICD9 codes "hyperlipidemia" and "pure hypercholesterolemia" to label FH patients, as there is not a distinct FH code. The drug data, on the other hand, appears to contain limited information that can benefit the model, presumably because both the case and control groups are using similar drugs.

Table 2. Important features for final FH model

	Feature name	Feature type	Associated class
1	Hyperlipidemia	Code	FH
2	Pure hypercholesterolemia	Code	FH
3	High cholesterol in LDL [Mass/volume] in Serum or Plasma by calculation	Lab	FH
4	Lipid panel	Notes	FH
5	Physical activity	Notes	FH
6	Specimen	Lab	FH
7	Exercise	Notes	FH
8	Culture - general	Lab	FH
9	Serum HDL/non-HDL cholesterol ratio measurement	Lab	FH
10	History of present illness	Notes	FH (very high/low)

Discussion

We have shown that even with only a small (N=71) set of gold-standard patients, we are able to attain well performing preliminary models for the identification of FH patients from the data contained in the EHR alone. This is impressive, given the nature of the data we can obtain for FH patients. Some will enter the healthcare system after having been treated elsewhere, making their lab measurements appear normal; some will have data in the system only for unrelated healthcare visits, where anything related to FH or cholesterol will not be recorded. As is always the case, many patients have missing data, and because Stanford is a tertiary care center, many of the patients in our cohort had only short periods of follow-up in our dataset.

It is important to note that our results only illustrate the potential performance of the model when deployed in a cardiovascular clinic setting. If deployed in a primary care setting, for instance, the resulting numbers of false positives and false negatives would change. Future work will investigate these differences. However, given the difficulty in distinguishing patients with high cholesterol in general from those with FH, this cohort provides a "worst case" scenario from which to start.

While the best performing model will, again, depend upon the intended setting and desired trade-off in sensitivity and specificity, both our default feature selection model and the no-feature-selection model performed well. However, we do hypothesize that limiting the feature set will make the final model more transportable to other hospital systems. In addition, we find that the inclusion of terms from clinical notes does not greatly improve performance in this use case, which is advantageous for deploying the model at hospitals where unstructured text data is not readily available for analysis.

While performance so far is promising, there are also additional model improvements that can be made. The largest jump in improvement will certainly come from an increase in the number of positive FH patients used in learning the model. In addition, our cohort of FH patients has so far been already medically identified, meaning that features such as "pure hypercholesterolemia" (which indicates an inherited form of high cholesterol) will give our model high performance; FH patients who have not been diagnosed will not have this code in their notes and will be harder to find with our current model. With larger cohorts of FH patients, more advanced feature engineering will also likely yield improvements in performance. For example, the actual lab measurement values, their changes over time, the range of measurements, and co-occurrence with medications could all be considered. We can additionally investigate the potential of tools such as topic models over terms from the notes to generate increased discrimination power from this feature type.

Another promising avenue for model improvement is the further adjustment of how the different feature types are combined. Recent work (17) has shown that building models on individual feature classes, and then creating meta-learners using these individual models, has resulted in improved performance in other diagnostic classification tasks.

This is certainly an area for future investigation. Incorporation of more advanced textual analysis, which will allow for identification of family history from the notes, could also provide a boost in classification performance.

Longer-term next steps will involve the deployment of the algorithm on additional EHR databases, followed by clinician review of identified cases to evaluate and further refine the algorithm. Once validated, we can deploy the algorithm on real-life EHR systems via the FIND FH program.

Conclusion

Patients with FH often are unaware of their condition, and physicians are currently poorly-trained to identify the disease. While in the best case an undiagnosed FH patient may be treated with a statin simply because he or she has high cholesterol, use of statins alone according to regular cholesterol care guidelines is often not enough for FH patients to attain healthy LDL levels. Their LDL management is much improved if they are correctly identified as having FH and managed accordingly.

The goal of our efforts to identify FH patients is somewhat a departure from typical phenotyping studies, in that rather than identifying a cohort of patients upon which to perform further analysis, we are attempting to alert a clinician about patients with FH who are unaware that they have it. The potential of our algorithm for benefitting patients, therefore, is very great – with a well-performing model and a functional outreach program, we can contact physicians whose patients may need an escalation in care.

Acknowledgements

KEN acknowledges funding from the Rhodes Trust and RCUK Digital Economy Programme grant number EP/G036861/1 (Centre for Doctoral Training in Healthcare Innovation). JK and NHS acknowledge funding from Amgen Inc, the American Heart Association and the Stanford Data Sciences Initiative.

References

1. Nordestgaard BG, Chapman MJ, Humphries SE, Ginsberg HN, Masana L, Descamps OS, et al. Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: Guidance for clinicians to prevent coronary heart disease. Eur Heart J. 2013;34(45):3478–90.
2. Foody JM. Familial hypercholesterolemia: An under-recognized but significant concern in cardiology practice. Clin Cardiol. 2014;37(2):119–25.
3. Peissig PL, Rasmussen L V., Berg RL, Linneman JG, McCarty C a., Waudby C, et al. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. J Am Med Informatics Assoc. 2012;19(2):225–34.
4. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc. 2013;21(2):221–30. Available from: http://jamia.bmj.com/content/early/2013/11/07/amiajnl-2013-001935.long.
5. Wang Z, Shah AD, Tate a. R, Denaxas S, Shawe-Taylor J, Hemingway H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. PLoS One. 2012;7(1).
6. Xu J, Rasmussen L V, Shaw PL, Jiang G, Kiefer RC, Mo H, et al. Review and evaluation of electronic health records-driven phenotype algorithm authoring tools for clinical and translational. J Am Med Informatics Assoc. 2015;1–12.
7. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio T a, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med. 2013 Oct [cited 2014 Apr 28];15(10):761–71. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3795928&tool=pmcentrez&rendertype=abstract.
8. PheKB. Phenotype Knowledge Base [Internet]. Available from: http://www.phekb.org.
9. Schulam P, Wigley F, Saria S. Clustering Longitudinal Clinical Marker Trajectories from Electronic Health Data: Applications to Phenotyping and Endotype Discovery. 2015.
10. Doshi-Velez F, Ge Y, Kohane I. Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis. Pediatrics. 2013;133(1):e54–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24323995.
11. Ross JC, Cho MH, Dy JG, Castaldi PJ. Dual Beta Process Priors for Latent Cluster Discovery in Chronic Obstructive Pulmonary Disease. :155–62.
12. Agarwal V, Lependu P, Barber R, Boland MR, Hripcsak G, Shah NH. Using narratives as a source to automatically learn phenotype models Department of Biomedical Informatics, Columbia University, New York , NY. DMMI. 2014. p. 1–7.
13. Agarwal V, Podchiyska T, Banda J, Goel V, Leung TI, Minty EP, et al. Learning statistical models of phenotypes using noisy labeled training data. [In Review]. 2015;1–19.
14. Halpern Y, Choi Y, Mmsc SH, Sontag D, Israel B, Medical D. Using Anchors to Estimate Clinical State without Labeled Data Anchor variable framework Observed and latent variables in the EMR. :606–15.
15. Carroll MD, Kit BK, Lacher D a, Yoon SS. Total and high-density lipoprotein cholesterol in adults: National Health and Nutrition Examination Survey, 2011-2012. NCHS Data Brief. 2013;2012(132):1–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24165064
16. Banda JM. Aphrodite [Internet]. 2015. Available from: https://github.com/OHDSI/Aphrodite.
17. Scheurwegs E, Luyckx K, Luyten L, Daelemans W, Van den Bulcke T. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. J Am Med Informatics Assoc. 2015;32(0):ocv115. Available from: http://jamia.oxfordjournals.org/lookup/doi/10.1093/jamia/ocv115.