PheValuator 2.0: Changes to Improve the Performance of the Phenotype Algorithm Evaluation Tool

Joel N. Swerdel, Martijn Schuemie, and Patrick B. Ryan

Background

A phenotype algorithm is the translation of the case definition of a health condition or phenotype into an executable algorithm that involves querying clinical data elements from a database. Developing and applying accurate phenotype algorithms is central to all observational analyses. PheValuator is an OHDSI tool aimed at estimating the performance characteristics, i.e., sensitivity, specificity, and positive and negative predictive value, of phenotype algorithms using the common data model (CDM).(1) PheValuator uses a semi-automated procedure and was developed to complement or even replace the traditional approach of algorithm validation, i.e., by expert assessment of subject records through chart review. The traditional method has several significant shortcomings including that it is time consuming, costly, and, in most cases, only provides positive predictive value (PPV). The last deficit is critical as sensitivity and specificity are the measures needed to correct misclassification bias in analytical results from observational studies. In the original PheValuator paper the authors assessed the performance of the tool itself against three algorithms for myocardial infarction evaluated in the literature using traditional algorithm validation. In that paper, it was shown that the tool provided a conservative estimate for PPV compared to previous results. The objective of this research is to detail a new method designed to improve the performance of the tool as compared to “gold standard” traditional evaluation techniques.

Methods

There are two main changes in the new method for PheValuator. The first change is to allow possible predictors used in the diagnostic predictive modeling step of the process to be from multiple time windows. In contrast, in the original design, all covariates used to fit the model were from a single time window spanning all time in the subject’s record. With this change, the tool accepts up to three time-windows for possible predictors relative to the start of the modeling time period. This is especially important for acute phenotypes, such as myocardial infarction, when the first days after the event may have many possible predictors that follow a specific time course. For acute phenotypes, we have used the time windows 0-30, 31-60, and 61-90 days after the start time to allow a modeling process more specific for acute events. Multiple time windows may also be used for chronic phenotypes, such as atrial fibrillation, where changes in patient treatment occur over a longer period. For chronic phenotypes, we have used time windows of 0-30, 31-60, and 61-9999 days after the start of the modeling period.

The second change is to the definition of the extremely specific, or xSpec, and extremely sensitive, or xSens, cohorts used to determine the cases and non-cases for the phenotype in the modeling step of the process. In the original method, the xSpec cohort was designed to provide subjects with a high probability of the phenotype by using an algorithm requiring multiple diagnosis codes of the phenotype in the subject’s record, for example requiring five separate mentions of ‘myocardial infarction’ diagnose codes such as ICD-10 code I21 for a ‘myocardial infarction’ phenotype. The non-cases were determined by only including subjects with no instances of the diagnosis code for the phenotype in their record achieved by excluding subjects in the xSens. While this ensured the inclusion of subjects with a high probability (cases) and low probability (non-cases) of the phenotype, it precluded using the diagnosis codes for the phenotype in the modeling process. If these codes were used in the model, the regularized regression model would not converge as there would be perfect separation between the cases and the non-cases. Not including the diagnosis codes for the phenotype may have been a limitation to the method as these codes, by definition, are important predictors for the phenotype. In order to potentially use the phenotype diagnosis codes in the model, we designed a new version of the xSpec that looks for the presence of diagnosis codes for the phenotype only in the time prior to the starting point for the model. The xSens cohort was redesigned similarly with a single code for the phenotype just prior to the starting time for the model. In this way we may include the diagnosis codes for the phenotype as possible predictors in our model by excluding the time window prior to the starting time.

To evaluate the performance of the new method, we compared the results from the new and original methods against results found from the literature using traditional validation of algorithms for five phenotypes (Table 1). We compared results for PPV only as these were the only performance characteristics provided by the prior studies. For this comparison we chose two acute phenotypes, myocardial infarction and ischemic stroke, and three chronic phenotypes, ankylosing spondylitis, atrial fibrillation, and Crohns disease. We performed these tests using data from three administrative claims databases, IBM® MarketScan® Commercial Database (CCAE), Multi-State Medicaid Database (MDCD), and Medicare Supplemental Database (MDCR) and one electronic health record database, Optum®'s longitudinal EHR repository (Optum EHR).

Table 1: Source and phenotype algorithms from prior validation studies.

Health Outcome of Interest	Author (Year)	Phenotype Algorithm
Myocardial Infarction	Cutrona (2013)(2)	International Classification of diseases, 9th edition (ICD-9) code for acute myocardial infarction (410.x0, 410.x1) in the principal or primary position on facility claims for hospitalizations
Ischemic Stroke	Thigpen (2015)(3)	ICD-9 codes for ischemic stroke (“with infarction” 433-434, 436) from hospital admissions
Ankylosing Spondylitis	Dubreille (2016)(4)	One or more ankylosing spondylitis diagnoses (Read code N100.00, “Ankylosing spondylitis”)
Atrial Fibrillation	Navar-Boggan (2015)(5)	ICD-9 code of 427.31 (“Atrial Fibrillation”) coded for 1 inpatient or 2 outpatient or emergency department visits
Crohns Disease	Ananthakrishnan (2013)(6)	At least 1 ICD-9 code for Crohns disease (555.x)

Results

The results for this study are shown in the Figure 1 below.

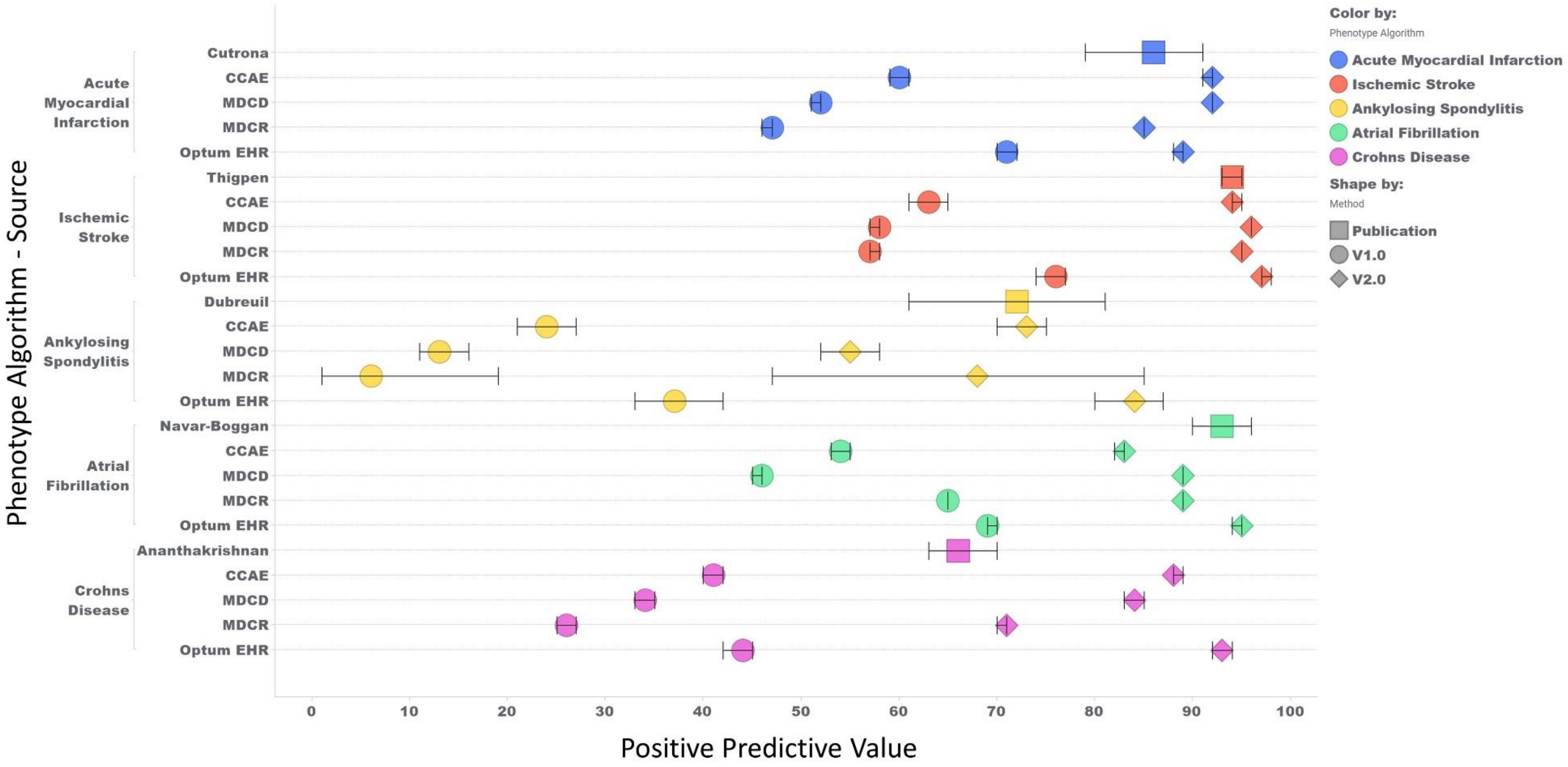

Figure 1. Comparison of results from the original method for PheValuator (V1.0) and the new method (V2.0) to results from prior publications of phenotype algorithm evaluation using traditional methods.

CCAE - IBM® MarketScan® Commercial Database, MDCD Multi-State Medicaid Database; MDCR - Medicare Supplemental Database; Optum EHR - Optum®'s longitudinal EHR repository; V1.0 – Original version of PheValuator, V2.0 – new version of PheValuator.

The results using the new process for PheValuator provide much closer agreement with the results from previously published validation studies compared to the original process. For example, in examining myocardial infarction, we found a range of PPVs in the original PheValuator method from 47% (MDCR) to 71% (Optum EHR). In the new method, the range of results for PPV was from 85% (MDCR) to 92% (CCAE, MDCD). The results from the new method were in close agreement with the results from Cutrona et al, where they determined PPV to be 86% (95% confidence interval 79-91%).(2)

Conclusion

We have developed a new method for the PheValuator tool to produce more accurate results for the evaluation of phenotype algorithms. We have found that the results from the new method are in closer agreement with prior results from published validation studies compared to results from the original method. With these enhancements, it may be possible to use these measures of performance to improve the validity of studies using observational data from health records.

References

1. Swerdel JN, Hripcsak G, Ryan PB. PheValuator: Development and evaluation of a phenotype algorithm evaluator. Journal of Biomedical Informatics. 2019;97:103258.
2. Cutrona SL, Toh S, Iyer A, Foy S, Daniel GW, Nair VP, et al. Validation of Acute Myocardial Infarction in the Food and Drug Administration's Mini-Sentinel program. Pharmacoepidemiol Drug Saf. 2013;22(1):40-54.
3. Thigpen JL, Dillon C, Forster KB, Henault L, Quinn EK, Tripodis Y, et al. Validity of international classification of disease codes to identify ischemic stroke and intracranial hemorrhage among individuals with associated diagnosis of atrial fibrillation. Circ Cardiovasc Qual Outcomes. 2015;8(1):8-14.
4. Dubreuil M, Peloquin C, Zhang Y, Choi HK, Inman RD, Neogi T. Validity of ankylosing spondylitis diagnoses in The Health Improvement Network. Pharmacoepidemiol Drug Saf. 2016;25(4):399-404.
5. Navar-Boggan AM, Rymer JA, Piccini JP, Shatila W, Ring L, Stafford JA, et al. Accuracy and validation of an automated electronic algorithm to identify patients with atrial fibrillation at risk for stroke. Am Heart J. 2015;169(1):39-44.e2.
6. Ananthakrishnan AN, Cai T, Savova G, Cheng SC, Chen P, Perez RG, et al. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. Inflamm Bowel Dis. 2013;19(7):1411-20.