

Development of a phenotype evaluation tool that emulates clinical expert chart review adjudication at scale

Joel N. Swerdel, PhD MS MPH^{1,2}, James P. Gilbert PhD^{1,2}, Judith A. Racoosin MD MPH², and George Hripcsak MD MS^{2,3}

¹Observational Health Data Analytics, Global Epidemiology, Janssen Research and Development, Titusville, NJ, USA

²Observational Health Data Sciences and Informatics (OHDSI), New York, NY, USA

³Department of Biomedical Informatics, Columbia University, New York, NY, USA

Background: Evaluation of phenotype algorithm performance is essential for interpreting observational research. Misclassification bias has been shown to effect estimates, outcome frequencies, and can impact the conclusions of effect estimation.[1] In practice, investigators seek estimates of sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) to understand how well a phenotype identifies true cases and true non-cases. One approach within OHDSI is PheValuator, which estimates phenotype performance through the use of diagnostic predictive modeling.[2] However, one limitation is that this process depends on complex LASSO regression models that can be difficult for end users to interpret. In addition, the resulting models are generally database dependent and do not transport readily across data systems. A more transparent process for evaluating phenotype performance would, therefore, be valuable.

A common practice for phenotype performance evaluation has been clinical expert adjudication of source medical records. However, medical record review is constrained by limited chart availability, substantial cost, and high execution time. These limitations motivate development of an alternative approach that more directly emulates expert adjudication while remaining scalable for large observational datasets.

Objective: The objective of this study was to develop a phenotype evaluation process that emulated expert adjudication by using rubrics and rule-based scoring procedures, and to assess whether this rubric-based approach produces phenotype performance estimates similar to those generated by the standard PheValuator process.

Methods: The rubric-based approach process began by selecting health conditions of interest. We then used the OpenAI o3 large language model (LLM) to generate a set of 7 rubrics for each condition, with each rubric designed to emphasize a distinct adjudication goal and to reflect alternative ways expert reviewers might weigh evidence. For example, one rubric was an early-course “peri-index management” rubric and focused on signals occurring around the index date (for example, from 14 days before through 30 days after index), while another was a longitudinal “post-index continuity of care” rubric designed to prioritize chronic follow-up patterns observed after index. Example rules for seizure are shown in Table 1. An example rule for seizure in the rubric for “peri-index management” may have a rule for “administration or dispensing of intravenous benzodiazepine or other rapid-acting anticonvulsant”, while the “post-index continuity of care” rubric may have a rule for “initiation of a new anti-epileptic medication different from prior agent”.

Within this analytic framework, the LLM assigned clinical rules within each rubric. The rules can be classified in 3 categories: 1) evidence rules (e.g., persons with seizure diagnosis and anti-epilepsy treatment) increase confidence in case classification, 2) alternative explanations (e.g., persons with migraine) decrease confidence in case classification, and 3) hard exclusion rules, (e.g., persons with electroconvulsive therapy procedure around the index date) where subjects should be disqualified as cases. The rules were translated into SQL code that was executed against cohorts of subjects drawn from the phenotype algorithms under evaluation. In principle, each rubric can be seen as a set of rules that can be applied to any individual on a given index date, giving an interpretable score that relates to how well the data confirms their correct inclusion in the cohort.

The LLM assigns a score to each rule based on the relative importance of the rule to the health condition. For each subject, rule scores are summed within each rubric to generate a rubric-specific score representing the likelihood of case status for the condition. Prespecified score thresholds established by the LLM classified subjects with scores at or above the upper threshold as cases; subjects with scores at or below the lower threshold as non-cases; and subjects with scores between these thresholds as indeterminate. In this analysis, subjects deemed indeterminate in the overall adjudication were treated as non-cases. The final decision for each subject is based on majority adjudication across the 7 rubrics.

To compare the rubric-based approach with PheValuator, we first used the PheValuator process to create a large random sample of subjects from each database. Each sampled subject was then adjudicated using both methods: PheValuator provided probabilistic values for case status, whereas the rubric process produced a case or non-case classification based on rubric scoring. We used these adjudications to classify phenotype performance for each cohort in each database. Sensitivity was calculated as true positives divided by true positives plus false negatives; PPV was calculated as true positives divided by true positives plus false positives. Analyses were conducted across 4 observational databases, and final sensitivity and PPV estimates were summarized as the mean across databases. We evaluated atrial fibrillation, seizure, urinary tract infection, and bladder cancer, focusing on simple phenotype cohorts defined by a single diagnosis-code inclusion criterion, and compared the resulting rubric-based estimates with those from the standard PheValuator process.

Results: Results are summarized in Figure 1. Across the 4 health conditions examined, the rubric-based approach produced estimates that were directionally similar to those from PheValuator on one of the 2 primary performance measures, while differing more substantially for the other. The seizure phenotype provides an illustrative example. For seizure, sensitivity was 87% using the rubric-based approach and 85% using PheValuator, indicating close agreement between methods in identifying true cases among all cases in the sample. In contrast, PPV for seizure was 59% with the rubric-based approach and 93% with PheValuator, reflecting a substantial difference in the proportion of phenotype-positive subjects classified as true cases. When subjects classified as indeterminate by the rubric process were instead counted as cases, the rubric-based PPV increased to 91% (data not shown), suggesting that treatment of indeterminate subjects account for an important share of the discrepancy between approaches. We also examined the contribution of negative scoring components within the rubric framework. When negative contributions from hard exclusions and alternative explanation rules were removed from the seizure scoring process, PPV increased to 70% (data not shown).

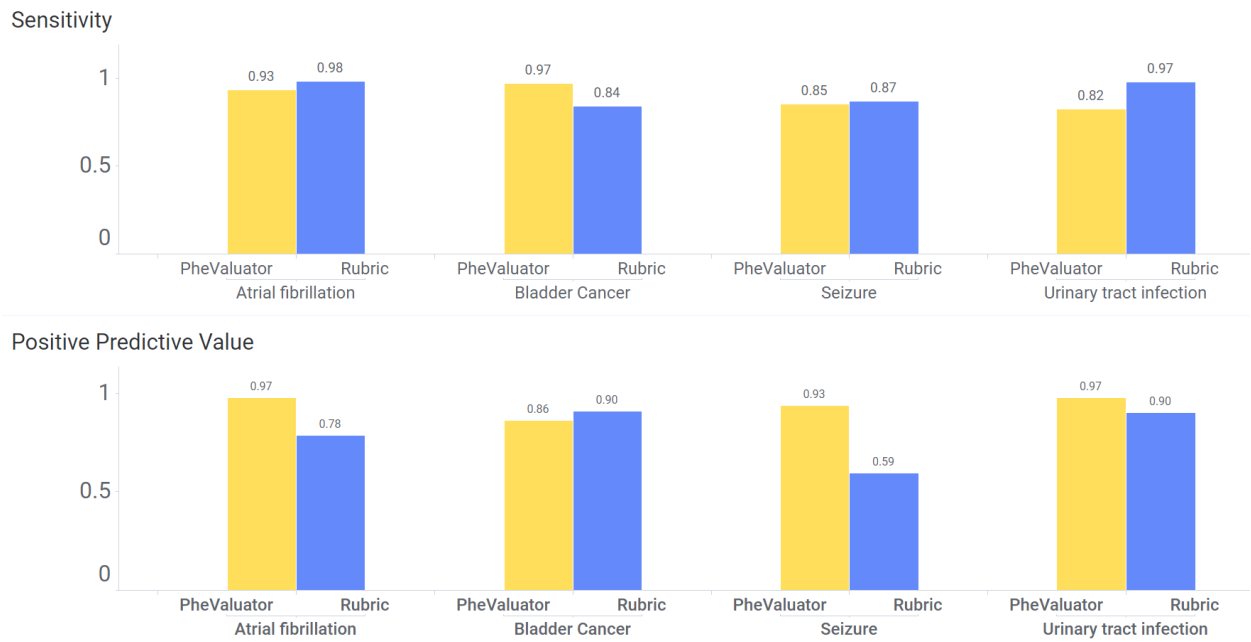
Conclusions: Across 4 health conditions, the rubric-based approach and PheValuator produced similar results for one of the 2 evaluated performance measures, supporting the feasibility of rubric-based adjudication as an alternative strategy for phenotype evaluation. The observed differences in the second measure appear to be influenced, at least in part, by how indeterminate subjects are handled and whether exclusionary and competing clinical explanations are explicitly incorporated into the scoring process.

Beyond empirical similarity, the rubric-based framework offers several practical advantages over PheValuator. First, it is database independent as a set of rubrics, representing standard clinical treatment, may be applied across any database. Second, it does not require a large number of subjects for model training and evaluation, making it potentially better suited for smaller databases where large subject populations may not be available. A third advantage arises from the transparency and flexibility of the rubric design itself. The rules and scoring system can be reviewed, modified, and refined by clinical or phenotype experts as needed, providing direct visibility into why a subject is classified as a case or non-case. Future research should compare rubric-based and clinical expert adjudication across a set of health conditions.

Table 1: Example rubric for seizure. A case required a score 6 and above; a non-case was applied to scores of 1 and below.

Type	Score	Criteria	Rationale
Hard exclusion	-100	Any diagnosis of eclampsia or severe preeclampsia within -1 to +7 of index	Obstetric seizures represent a distinct clinical entity; label validity cannot be confirmed for primary seizure disorder.
	-100	Electroconvulsive therapy procedure within -7 to +7 of index	ECT induces controlled seizures; any seizure diagnosis in this context is procedure-related, not spontaneous epilepsy.
Alternative explanations	-4	Diagnosis of psychogenic non-epileptic seizure or conversion disorder within -30 to +60 of index	PNES often miscoded as epileptic seizure.
	-3	Diagnosis of hypoglycemia WITH coma within -5 to +5 of index	Hypoglycemic seizures may mimic epilepsy.
	-3	Transient ischemic attack or acute cerebrovascular accident diagnosis within -5 to +5 of index	Cerebrovascular events are common mimics.
	-2	Cardiac arrhythmia WITH syncope diagnosis within -5 to +5 of index	Need to reduce confidence when competing cardiogenic cause present.
Evidence rules	2	(a) ONE seizure diagnosis within an inpatient OR emergency department visit within -1 to +1 of index OR (b) TWO outpatient seizure diagnoses separated by 1-60 days both within -60 to +60 of index	
	4	Diagnosis of status epilepticus within +1 to +30 of index	
	2	Diagnosis of aspiration pneumonia within +1 to +14 of index	
	3	Dispensing of any anti-seizure medication within -30 to +60 of index	
	3	Administration or procedure code for parenteral benzodiazepine (e.g., diazepam, lorazepam) within -1 to +1 of index	
	2	Outpatient visit with neurologist specialty within +7 to +60 of index	
	2	EEG procedure code within -1 to +7 of index	
	1	Diagnosis of closed fracture of upper limb, lower limb, or facial bones within +1 to +14 of index	

Figure 1: Comparison of PheValuator and Rubric analysis of sensitivity and positive predictive value. The data represents means across four observational databases for four health conditions.



1. Lash TL, Fox MP, Fink AK. Applying Quantitative Bias Analysis to Epidemiologic Data: Springer New York, NY; 2009.
2. Swerdel JN, Schuemie M, Murray G, Ryan PB. PheValuator 2.0: Methodological improvements for the PheValuator approach to semi-automated phenotype algorithm evaluation. J Biomed Inform. 2022:104177.