Evaluating a phenotype using PheValuator

Patrick Ryan, PhD

On behalf of Joel Swerdel, PhD

Janssen Research and Development

Agenda

- What is a phenotype and why do we need them?
- Why do we need a phenotype evaluator?
- Development of the evaluator
- Results from the evaluation

Case Definitions and Phenotyping Algorithms

- "A case definition describes characteristics that a patient must possess to have a disease from a clinical perspective."
- "An EHR phenotyping algorithm is the translation of the case definition into an executable algorithm that involves querying clinical data elements from the EHR."

J Am [illegible] PMCID: [illegible]
Published online 2013 Jul 9. doi: 10.1136/amiajnl-2013-001930 PMID: 23837993

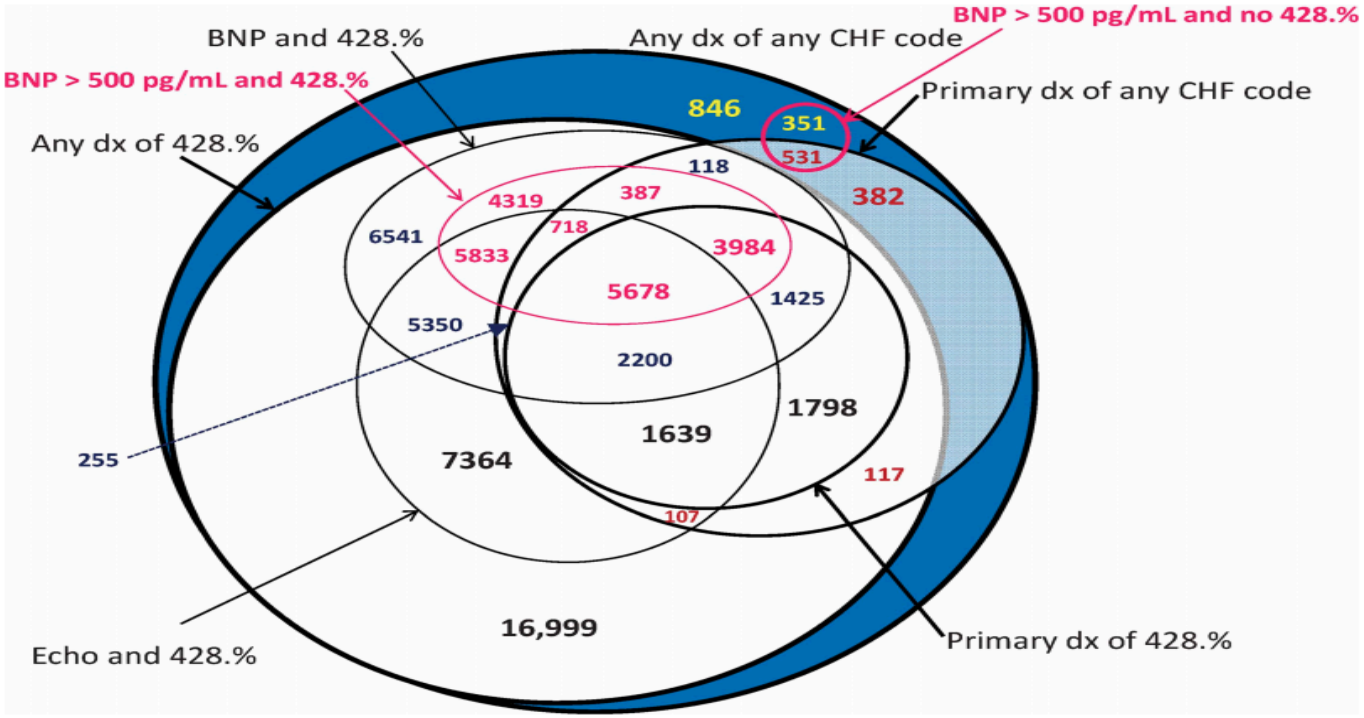
A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury

[illegible][1,2] [illegible][3] [illegible][4,5] [illegible][1] [illegible][1] [illegible][3] Kevin Bruce,[3] Stephanie Johnson,[6] Jayant Talwalkar,[6] Yufeng Shen,[1,7] Steve Ellis,[5,8] Iftikhar Kullo,[6] [illegible][3] [illegible], Erwin Bottinger,[5,9,10] George Hripcsak,[1] and Chunhua Weng[1]

Evaluating cohort definitions

- How do we know if our cohort definition is any “good”?
- What is our goal for a cohort definition’s performance for a given use case?
- How do we know if our cohort definition is generalizable across sites?

Example: Evaluating CHF definitions



Rosenman et al. Database queries for hospitalizations for acute congestive heart failure: flexible methods and validation based on set theory. *J Am Med Inform Assoc*. 2014 Mar-Apr;21(2):345-52.

Example: Evaluating CHF definitions

Table 3 Results for the 10 congestive heart failure (CHF) phenotype queries

Criteria to combine Venn diagram zones	N in query	Sensitivity (%)	Sensitivity, SE (%)	PPV (%)	PPV, SE (%)
Any CHF	66 942	94.3	1.3	42.8	1.5
Any dx of 428	64 832	90.9	1.3	42.5	1.5
Any dx of CHF and BNP >500 pg/mL	21 801	50.8	1.8	70.7	2.5
1^0 dx of any CHF	19 339	54.8	1.9	86.0	2.2
1^0 dx of 428	16 724	47.6	1.7	86.3	2.5
1^0 dx of any CHF and BNP >500 pg/mL	11 298	33.5	1.3	90.0	2.1
1^0 dx of 428 and BNP >500 pg/mL	9662	28.8	1.1	90.4	2.4
1^0 dx of 428 and BNP >500 pg/mL and echocardiogram	5678	16.2	0.8	86.6	3.5
1^0 dx of any CHF or BNP >500 pg/mL	29 587	71.4	2.1	73.3	2.2
1^0 dx of 428 or BNP >500 pg/mL	28 863	69.6	2.1	73.2	2.2
High BNP, no ICD-9 diagnosis for CHF					
Zone X: no ICD-9 dx of 428, but BNP >500 pg/mL	12 149	N/A	N/A	14.3	3.5

BNP, B-natriuretic peptide; PPV, positive predictive value.

Ground Truth?

- To measure performance we need an outcome such as ‘case’ and ‘not a case’
- This determination is typically based on expert review of available data (e.g., sometimes will have notes etc that are not part of definition)
- The review process may include some heuristic guidance to ensure consistency amongst reviewers + Cohen’s Kappa
- Some newer research into automated ways to assess true cases (e.g., cohort characteristics)

Performance Metrics

- PPV is currently the primary metric obtained through manual review
- Sensitivity is sometimes determined when there are sufficient resources or when the incidence rate is reasonably high

Graham et al's discussion of outcome 'validation'

The codes defining ischemic stroke have a positive predictive value (PPV) of 88% to 95%.[18–20] Major bleeding was defined as a fatal bleeding event, a hospitalized bleeding event requiring transfusion,
(ie, intracra
retroperiton
Intracranial
atraumatic
for hemorr
validated. V
a bleeding e
related. The
86% to 88%
a PPV of 8
and 97% in

Rewriting to state our knowledge about data quality:

"Somewhere between 1-in-10 and 1-in-20 patients who have one of the diagnosis codes for ischemic stroke DO NOT actually have ischemic stroke."

"We DO NOT know how many people who don't have the stroke diagnosis codes actually DO have ischemic strokes (e.g. missing data, miscoding, censoring – death before heath service utilization), or whether these false negatives represent a differential bias."

Case Definition – Myocardial Infarction

Published by Oxford University Press on behalf of the International Epidemiological Association
© The Author 2010; all rights reserved. Advance Access publication 5 October 2010
International Journal of Epidemiology 2011;40:139–146
doi:10.1093/ije/dyq165

CARDIOVASCULAR DISEASE

World Health Organization definition of myocardial infarction: 2008–09 revision

Shanthi Mendis,[1]* Kristian Thygesen,[2] Kari Kuulasmaa,[3] Simona Giampaoli,[4] Markku Mähönen,[3] Kathleen Ngu Blackett,[5] Liu Lisheng[6] and Writing group on behalf of the participating experts of the WHO consultation for revision of WHO definition of myocardial infarction

- "MI is defined by **the demonstration of myocardial cell necrosis due to significant and sustained ischaemia.**"
- (i) ECG showing pathological Q waves and/or ST segment elevation or depression;
- (ii) history of typical or atypical angina pectoris, together with changes on the ECG and elevated enzymes;
- (iii) history of typical angina pectoris and elevated enzymes with no changes on the ECG or not available

Phenotyping Algorithm

Abstract

Purpose—To validate an algorithm based upon International Classification of Diseases, 9th revision, Clinical Modification (ICD-9-CM) codes for acute myocardial infarction (AMI) documented within the Mini-Sentinel Distributed Database (MSDD).

Methods—**Using an ICD-9-CM-based algorithm (hospitalized patients with 410.x0 or 410.x1 in primary position),** we identified a random sample of potential cases of AMI in 2009 from 4 Data Partners participating in the Mini-Sentinel Program. Cardiologist reviewers used information abstracted from hospital records to assess the likelihood of an AMI diagnosis based on criteria from the joint European Society of Cardiology and American College of Cardiology Global Task Force. Positive predictive values (PPVs) of the ICD-9-based algorithm were calculated.

Results—Of the 153 potential cases of AMI identified, hospital records for 143 (93%) were retrieved and abstracted. Overall, the PPV was 86.0% (95% confidence interval; 79.2%, 91.2%). PPVs ranged from 76.3% to 94.3% across the 4 Data Partners.

Conclusions—The overall PPV of potential AMI cases, as identified using an ICD-9-CM-based algorithm, may be acceptable for safety surveillance; however, PPVs do vary across Data Partners. This validation effort provides a contemporary estimate of the reliability of this algorithm for use in future surveillance efforts conducted using the FDA's MSDD.

What is a phenotype and why do we need them

- Tendency to equate the case definition with the phenotype algorithm (or the cohort definition) – the algorithm is the coded *approximation* of the case definition.
- Case definitions must be translated into algorithms for working with observational datasets
- But many properties of case definitions are lost in an algorithm causing imprecision when using an algorithm
- How much imprecision? → Need for validation

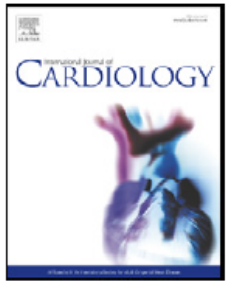
Validating Algorithms

- Many research studies have attempted to validate algorithms

Contents lists available at ScienceDirect

International Journal of Cardiology

journal homepage: www.elsevier.com/locate/ijcard



Review

Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: A systematic review and recommendations

Bruna Rubbo [a,*], Natalie K. Fitzpatrick [a], Spiros Denaxas [a], Marina Daskalopoulou [b], Ning Yu [a], Riyaz S. Patel [a,c], UK Biobank Follow-up and Outcomes Working Group, Harry Hemingway [a]

- Examined 33 studies
- Found significant heterogeneity in algorithms used, validation methods, and results

Validating an Algorithm

		Truth	
		Positive	**Negative**
Test	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Test – Comes from the algorithm/cohort definition

Truth – Some form of "gold standard" reference

Ex.: True Positive (TP) – Test and Truth agree Positive

For a complete validation of the algorithm we need:

1) Sensitivity: TP / (TP + FN)
2) Specificity: TN / (TN + FP)
3) Positive Predictive Value: TP / (TP + FP)

Evaluating Performance of Algorithm - Examples

Abstract

Purpose—To validate an algorithm based upon International Classification of Diseases, 9th revision, Clinical Modification (ICD-9-CM) codes for acute myocardial infarction (AMI) documented within the Mini-Sentinel Distributed Database (MSDD).

Methods—**Using an ICD-9-CM-based algorithm (hospitalized patients with 410.x0 or 410.x1 in primary position),** we identified a random sample of potential cases of AMI in 2009 from 4 Data Partners participating in the Mini-Sentinel Program. **Cardiologist reviewers used information abstracted from hospital records to assess the likelihood of an AMI diagnosis based on criteria from the joint European Society of Cardiology and American College of Cardiology Global Task Force.** Positive predictive values (PPVs) of the ICD-9-based algorithm were calculated.

Results—Of the 153 potential cases of AMI identified, hospital records for 143 (93%) were retrieved and abstracted. **Overall, the PPV was 86.0% (95% confidence interval; 79.2%, 91.2%).** PPVs ranged from 76.3% to 94.3% across the 4 Data Partners.

Conclusions—The overall PPV of potential AMI cases, as identified using an ICD-9-CM-based algorithm, may be acceptable for safety surveillance; however, PPVs do vary across Data Partners. This validation effort provides a contemporary estimate of the reliability of this algorithm for use in future surveillance efforts conducted using the FDA's MSDD.

Evaluating Performance of Algorithm - Examples

PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2009; **18**: 1064–1071
Published online 28 August 2009 in Wiley InterScience (www.interscience.wiley.com) **DOI**: 10.1002/pds.1821

ORIGINAL REPORT

SUMMARY

Purpose Studies of non-steroidal anti-inflammatory drugs (NSAIDs) and cardiovascular events using administrative data require identification of incident acute myocardial infarctions (AMIs) and information on whether confounders differ by NSAID status.

Methods **We identified patients with a first AMI hospitalization from Tennessee Medicaid files as those with primary ICD-9 discharge diagnosis 410.x and hospitalization stay of>2 calendar days.** Eligible persons were non-institutionalized, aged 50–84 years between 1999–2004, had continuous enrollment and no AMI, stroke, or non-cardiovascular serious medical illness in the prior year. Of 5534 patients with a potential first AMI, a systematic sample (n¼350) was selected for review. **Using defined criteria, we classified events using chest pain history, EKG, and cardiac enzymes, and calculated the positive predictive value (PPV) for definite or probable AMI.**

Results 337 of 350 (96.3%) charts were abstracted. 300 (89.1%), 5 (1.5%), and 32 (9.5%) events were categorized as definite, probable, and no AMI, respectively. **PPV for any definite or probable AMI was 92.8% (95% CI 89.6–95.2); for an AMI without an event in the past year 91.7% (95% CI 88.3–94.2), and for an incident AMI was 72.7% (95% CI 67.7–77.2).** Age-adjusted prevalence of current smoking (46.4% vs. 39.1%, p¼0.35) and aspirin use (36.9% vs. 35.9%, p¼0.90) was similar among NSAID users and non-users

Conclusions ICD-9 code 410.x had high predictive value for identifying AMI. Among those with AMI, smoking and aspirin use was similar in NSAID exposure groups, suggesting these factors will not confound the relationship between NSAIDs and cardiovascular outcomes.

Evaluating Performance of Algorithm - Examples

Yonsei Medical Journal
Vol. 41 No. 5 pp. 570—576 2000

Abstract

We attempted to assess the accuracy of the International Classification of Diseases (ICD) codes for myocardial infarction (MI) in medical insurance claims, and to investigate the reasons for any inaccuracy. This study was designed as a preliminary study to establish a surveillance system for cardiovascular diseases in Korea. A sample of 258 male patients who were diagnosed with MI from 1993 to 1997 was selected from the Korea Medical Insurance Corporation cohort (KMIC cohort: 183,461 people). The registered medical record administrators were trained in the survey technique, and gathered data by investigating the medical records of the study subjects from March 1999 to May 1999. **The definition of MI for this study included symptoms pursuant to the diagnostic criteria of chest pain, electrocardiogram (ECG) findings, cardiac enzyme and results of coronary angiography or nuclear scan.** We asked the record administrators for the reasons of incorrectness for cases where the final diagnosis was 'not MI'. **The accuracy rate of the ICD codes for MI in medical insurance claims was 76.0% (196 cases) of the study sample,** and 3.9% (ten cases) of the medical records were not available due to hospital closures, non-computerization or missing information. Nineteen cases (7.4%) were classified as insufficient due to insufficient records of chest pain, ECG findings, or cardiac enzymes. The major reason of inaccuracy in the disease code for MI in medical insurance claims was 'to meet the review criteria of medical insurance benefits (45.5%)'. The department responsible for the inaccuracy was the department of inspection for medical insurance benefit of the hospitals.

Evaluating Performance of Algorithm

Author (year; country)	*n*	Cross-referencing elements				PPV% (95%CI)
		Markers*	ECG	Symptom	Others*	
Secondary care EHR vs. chart review						
Gronski *et al.* (2012; USA)	294				●	20.0 (16.4-25.7)
Roger *et al.* (2002; USA)[a]	4061	●	●	●	●	40 (38.5-41.5)*
Kimm *et al.* (2012; South Korea)[d]	78	●				73.1 (62-82)*
[illegible]	[illegible]	●	●	●	●	[illegible]
Ryu *et al.* (2000; South Korea)	258	●	●	●	●	76 (70.4-80.8)*
[illegible]	[illegible]	●	●		●	[illegible]
Joensen *et al.* (2008; Denmark)	1072	●	●	●	●	81.9 (79.5-84.2)
Metcalfe *et al.* (2013; Canada)	169	●			●	82.8 (76-88)*
[illegible]	[illegible]					[illegible]
Cutrona *et al.* (2012; USA)	143	●	●	●	●	86.0 (79.2-91.2)
[illegible]	[illegible]	●	●	●	●	[illegible]
Whal *et al.* (2010; USA)	200	●			●	88.4 (83.2-92.5)
[illegible]	[illegible]	●	●	●	●	[illegible]
Choma *et al.* (2009; USA)	337	●	●	●	●	92.8 (89.6-95.2)
[illegible]	[illegible]	●	●	●		[illegible]
Barchielli *et al.* (2010; Italy)	372	●	●	●	●	94.6 (92.3-96.9)
Hammar *et al.* (2001; Sweden)	713	●	●		●	95 (93.1-96.3)*
Varas-Lorenzo *et al.* (2008; Canada)	193	●	●	●	●	95 (91-98)
Harriss *et al.* (2011; Australia)	202	●	●		●	95.5 (91.7-97.6)
Quan *et al.* (2008; Canada)	385	●			●	95.9 (93.4-97.4)*
Yeh *et al.* (2010; USA)	640	●	●			96.7 (95.0-97.8)*
Linnersjo *et al.* (2000; Sweden)	2101	●	●	●	●	98 (97.2-98.5)*
Coloma *et al.* (2013; Danish data)	148	●	●	●	●	100.0 (100-100)

Evaluating Performance of Algorithm

- Conclusion – for MI → no "gold standard" algorithm available
- Process is very costly and time consuming
- Small sample sizes → wide variation in estimates with wide confidence intervals

- In 33 studies "validating" algorithms, all reported PPV but:
  - Only 11 reported sensitivity
  - Only 5 reported specificity
  - **Is this really validation?**

The Value of Positive Predictive Value

- PPV is almost always reported in validation studies – easiest to assess
- Sensitivity and Specificity much less frequently reported
  - High cost and time to evaluate
- BUT – sensitivity and specificity are the actual characteristics of the test
  - PPV is a function of sensitivity, specificity and **prevalence** of Heath Outcome of Interest (HOI)

PPV Example – 1 Test, 2 Populations

Test Characteristics:

Sensitivity = 75%

Specificity = 99.9%

Population = 10,000

Prevalence = 1%		Truth	
		Positive	Negative
Test	Positive	75	10
	Negative	25	9890
	Total	100	9900

PPV =
75 / (75 + 10) =
88%

Prevalence = 5%		Truth	
		Positive	Negative
Test	Positive	375	10
	Negative	125	9490
		500	9500

PPV =
375 / (375 + 10) =
97%

PPV Example – 1 Population, 2 Tests

PPV = 90%

Population = 10,000

Prevalence = 5%		Truth	
		Positive	Negative
Test	Positive	90	10
	Negative	410	9490
	Total	500	9500

PPV = 90/(90+10) = 90%

Sens = 90/500 = 18%

Spec = 9490/9500 = 99.9%

Prevalence = 5%		Truth	
		Positive	Negative
Test	Positive	360	40
	Negative	140	9460
		500	9500

PPV = 360/(360+40) = 90%

Sens = 360/500 = 72%

Spec = 9460/9500 = 99.6%

Living with Algorithms

- Algorithms are used in most research with observational data
- Many ways to define an algorithm for any health outcome
- Each definition will have its own performance characteristics
  - Need to validate the algorithm to understand these characteristics
- Validation of an algorithm to be used in an observational dataset through chart review is likely not possible
  - Costly
  - Time consuming
  - Data is usually not available
- What do we really get from the from published phenotype algorithm validations?

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

PheValuator: Development and evaluation of a phenotype algorithm evaluator

Joel N. Swerdel[a,b,*], George Hripcsak[b,c], Patrick B. Ryan[a,b,c]

[a] Janssen Research & Development, 920 Route 202, Raritan, NJ 08869, USA
[b] OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), 622 West 168th Street, PH-20, New York, NY 10032, USA
[c] Columbia University, 622 West 168th Street, PH20, New York, NY 10032, USA

ARTICLE INFO

Keywords:
Phenotype algorithms
Validation
Diagnostic predictive modeling

ABSTRACT

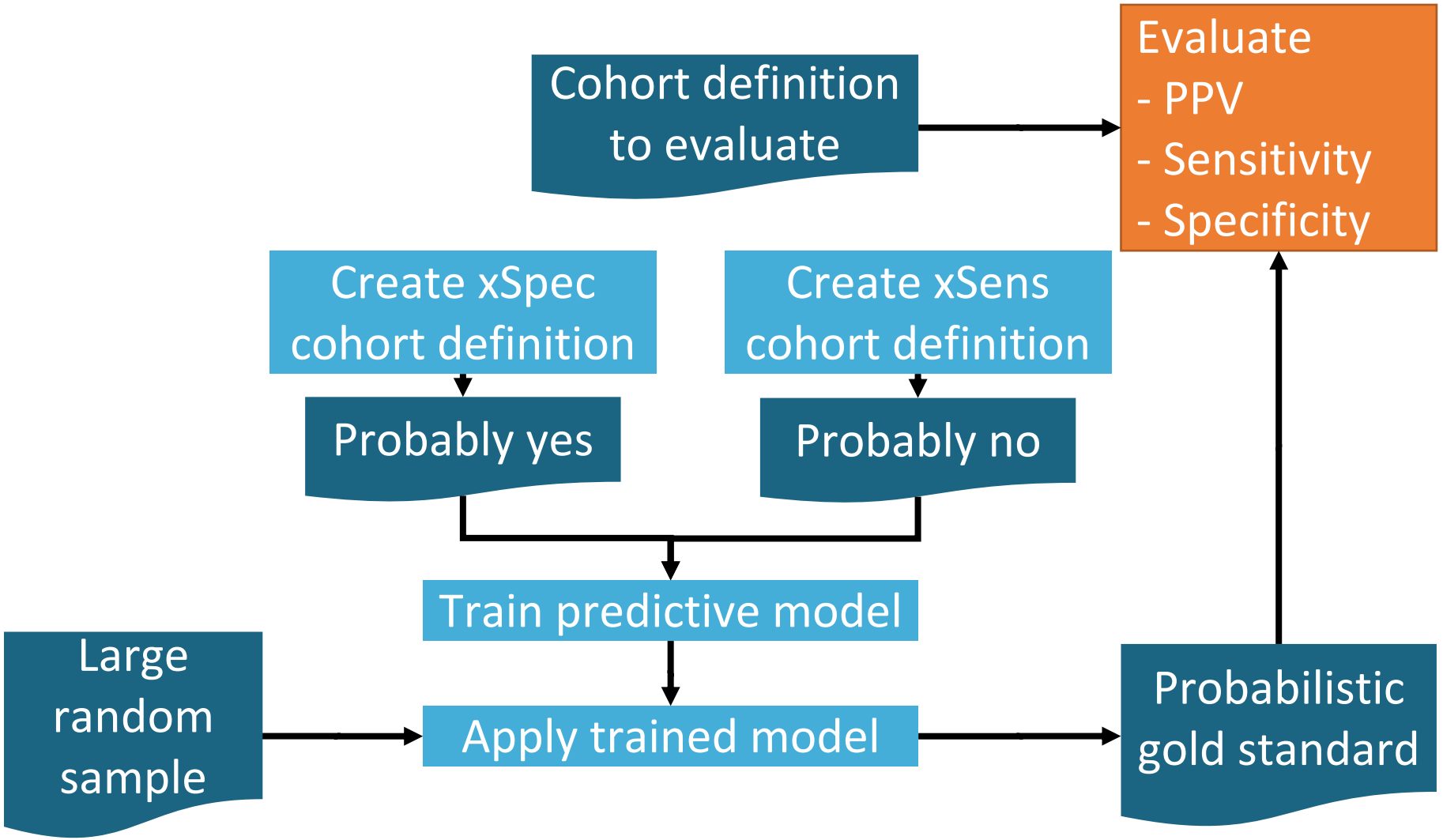
Background: The primary approach for defining disease in observational healthcare databases is to construct phenotype algorithms (PAs), rule-based heuristics predicated on the presence, absence, and temporal logic of clinical observations. However, a complete evaluation of PAs, i.e., determining sensitivity, specificity, and positive predictive value (PPV), is rarely performed. In this study, we propose a tool (PheValuator) to efficiently estimate a complete PA evaluation.

Methods: We used 4 administrative claims datasets: OptumInsight's de-identified Clinformatics™ Datamart (Eden Prairie,MN); IBM MarketScan Multi-State Medicaid); IBM MarketScan Medicare Supplemental Beneficiaries; and IBM MarketScan Commercial Claims and Encounters from 2000 to 2017. Using PheValuator involves (1) creating a diagnostic predictive model for the phenotype, (2) applying the model to a large set of randomly selected subjects, and (3) comparing each subject's predicted probability for the phenotype to inclusion/exclusion in PAs. We used the predictions as a 'probabilistic gold standard' measure to classify positive/negative cases. We examined 4 phenotypes: myocardial infarction, cerebral infarction, chronic kidney disease, and atrial fibrillation. We examined several PAs for each phenotype including 1-time (1X) occurrence of the diagnosis code in the subject's record and 1-time occurrence of the diagnosis in an inpatient setting with the diagnosis code as the primary reason for admission (1X-IP-1stPos).

Results: Across phenotypes, the 1X PA showed the highest sensitivity/lowest PPV among all PAs. 1X-IP-1stPos yielded the highest PPV/lowest sensitivity. Specificity was very high across algorithms. We found similar results between algorithms across datasets.

Conclusion: PheValuator appears to show promise as a tool to estimate PA performance characteristics.

Clinical Validity



Validating Algorithms in Observational Data

		Truth	
		Positive	**Negative**
Test	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Test – Comes from the algorithm/cohort definition

Truth – Some form of "gold standard" reference

Possible alternative for finding the "Truth"

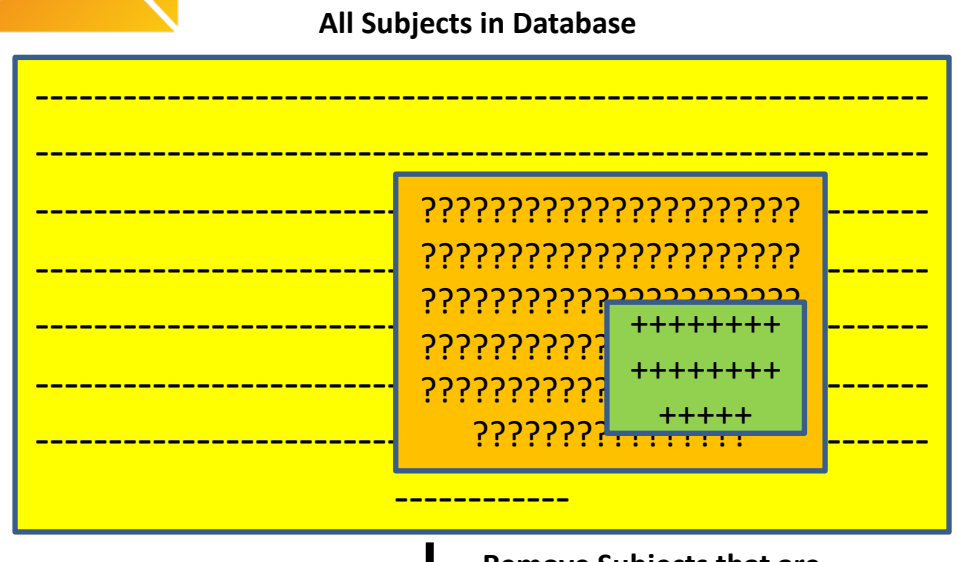
Diagnostic Predictive Models

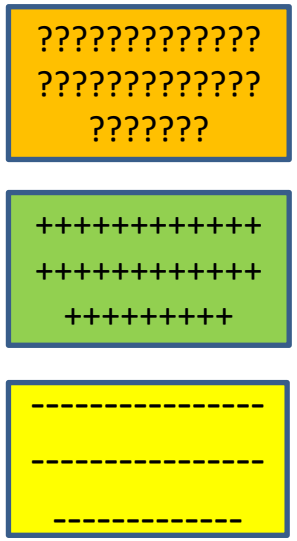
Prediction models used to estimate the probability of having a particular disease or outcome.

Diagnostic Predictive Models

- A predictive model is developed using a set of labeled data where the label represents the presence or absence of the HOI for each subject in the dataset
- The more accurate the labels, the more precisely the model will be able to determine the predictors that discriminate between those with the HOI and those without
- We wanted to only include subjects for which we were very likely to know the correct HOI label (presence/absence).

Creating the Population for the Model





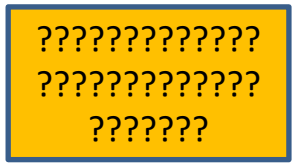
Possibly a case for the HOI → Find subjects using a very **sensitive** phenotype algorithm

Very likely a case for the HOI → Find subjects using a very **specific** phenotype algorithm

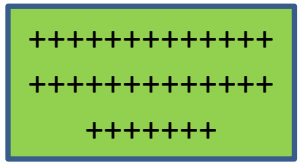
Very likely not a case for the HOI → Find subjects **NOT** in a very **sensitive** phenotype algorithm

Models with correct labels will produce models with better performance characteristics

Algorithms for Labeling the Subjects

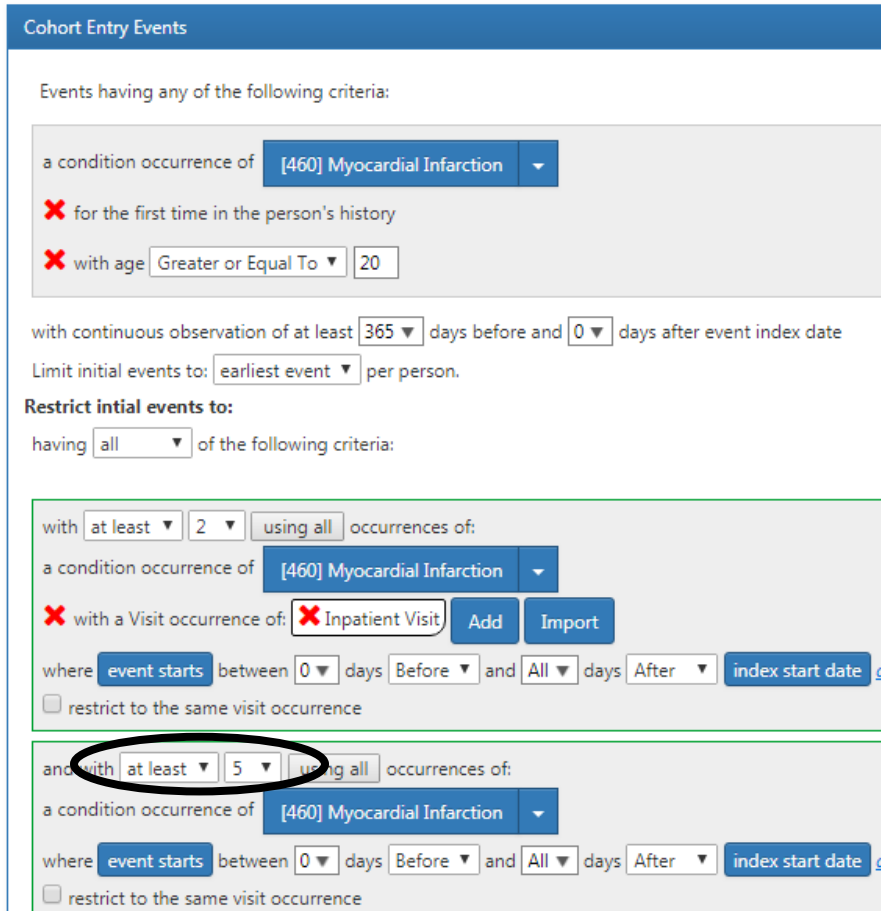


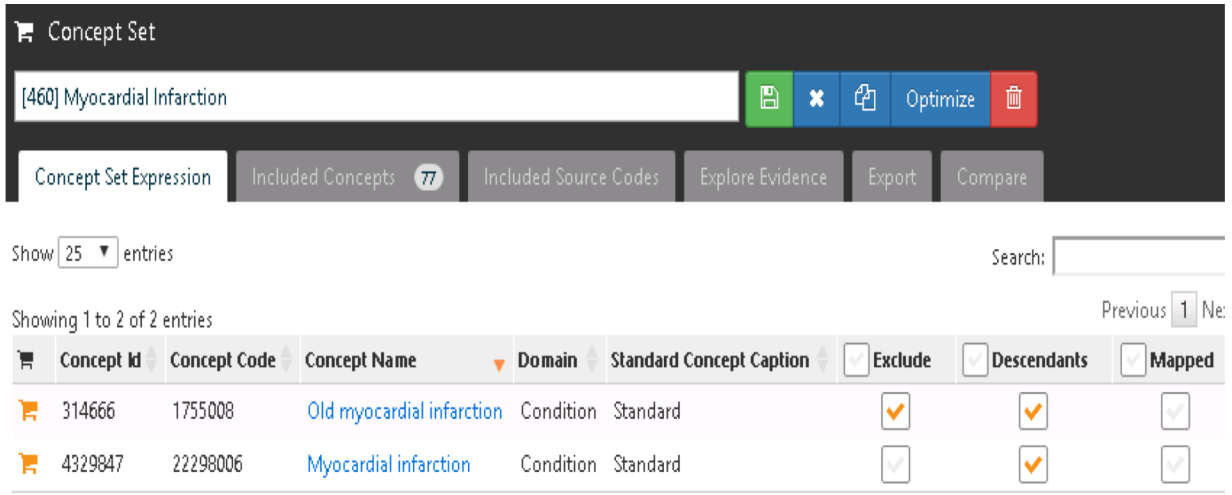
Possibly a case for the HOI → Find subjects using a very **sensitive** phenotype algorithm → Use a phenotype algorithm requiring subjects to have at **least one occurrence of a condition code (“**≥1X HOI”) for the HOI



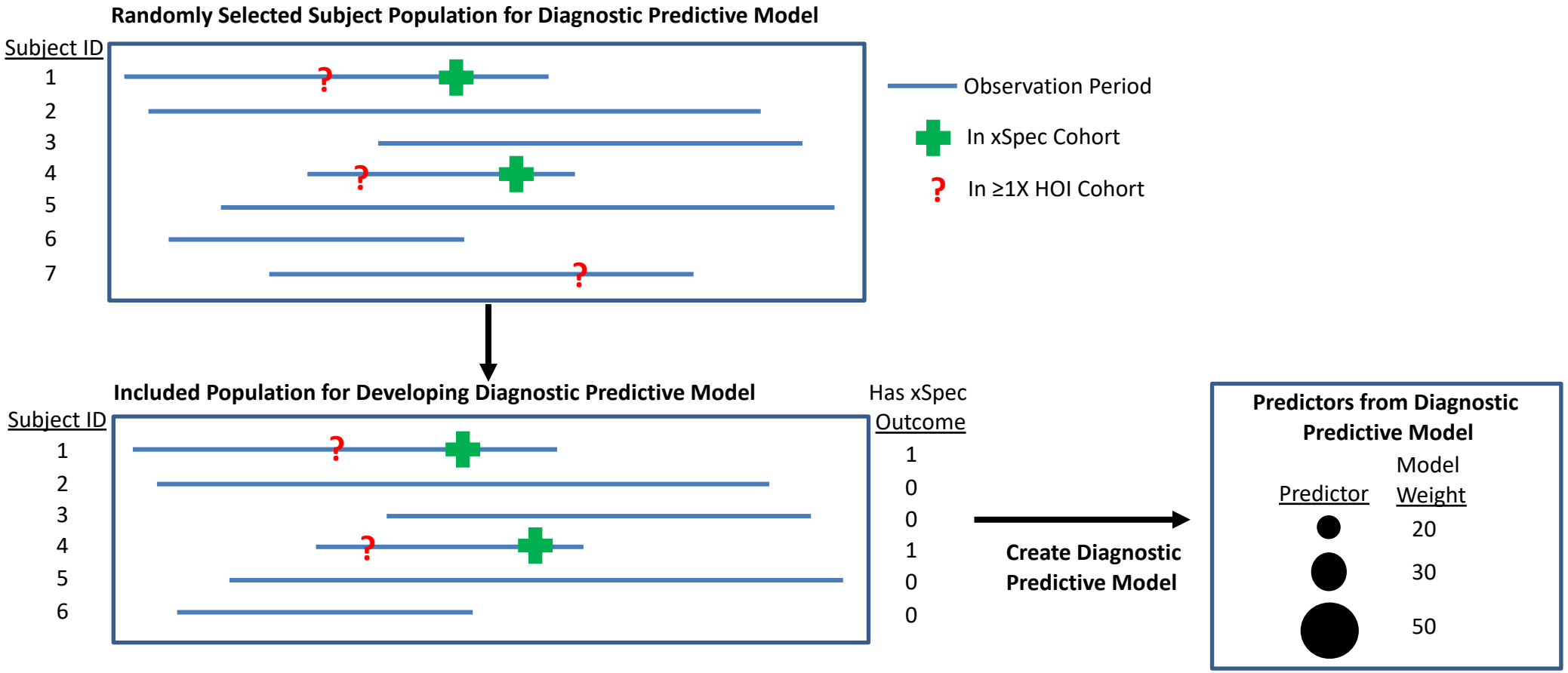
Very likely a case for the HOI → Find subjects using a very **specific** phenotype algorithm → Use a phenotype algorithm requiring subjects to **have many occurrences of a condition code** for the HOI
→ an **extremely specific (xSpec)** phenotype algorithm

Extremely specific (xSpec) cohort





Creating the Diagnostic Predictive Model

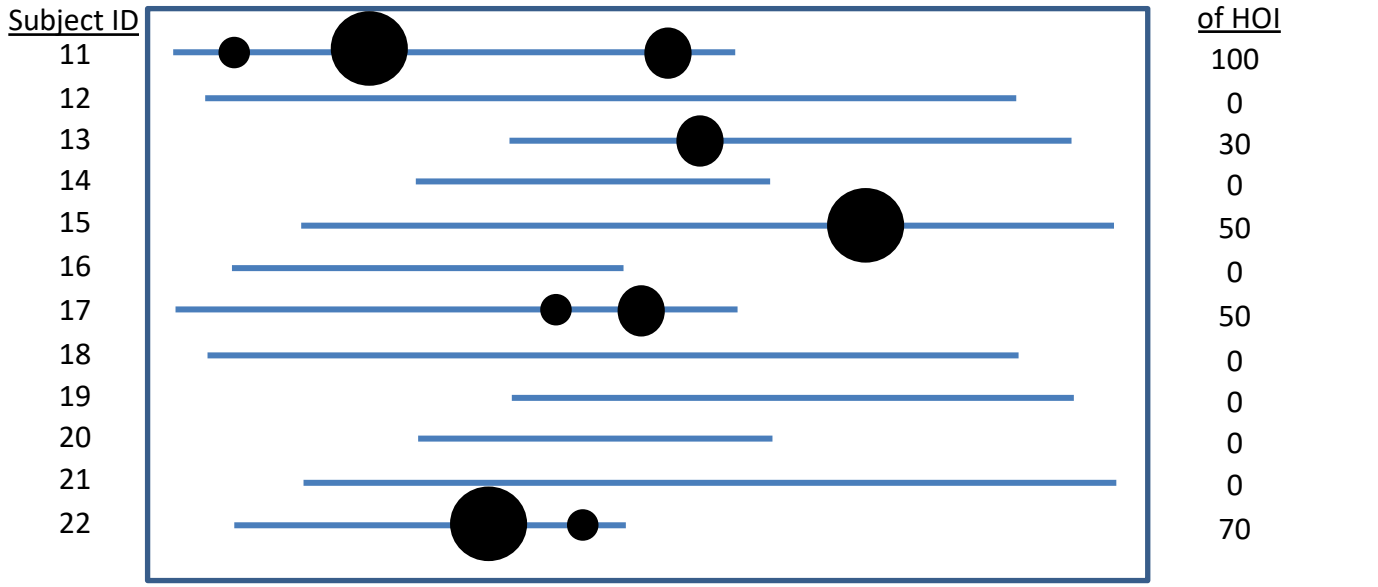


Developing an Evaluation Cohort

Predictors from Diagnostic Predictive Model

Predictor	Model Weight
●	20
●	30
●	50

Randomly Selected Subject Population For Evaluation Cohort

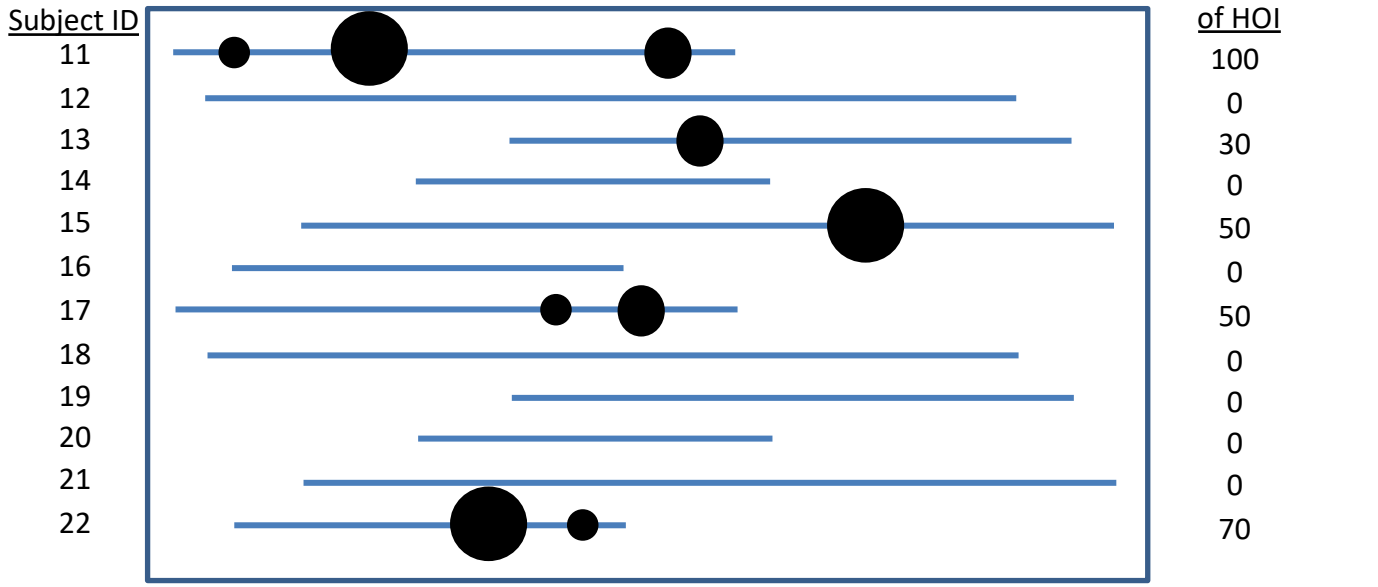


Developing an Evaluation Cohort

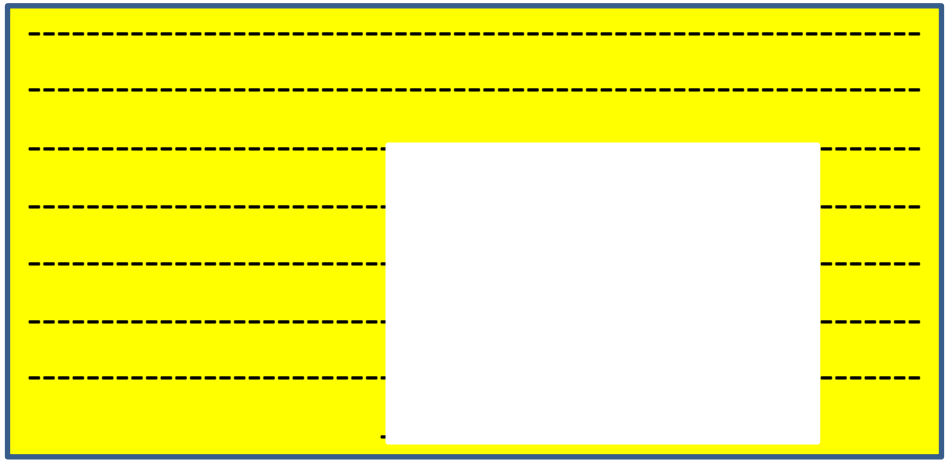
Predictors from Diagnostic Predictive Model

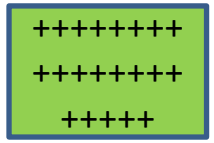
Predictor	Model Weight
●	20
●	30
●	50

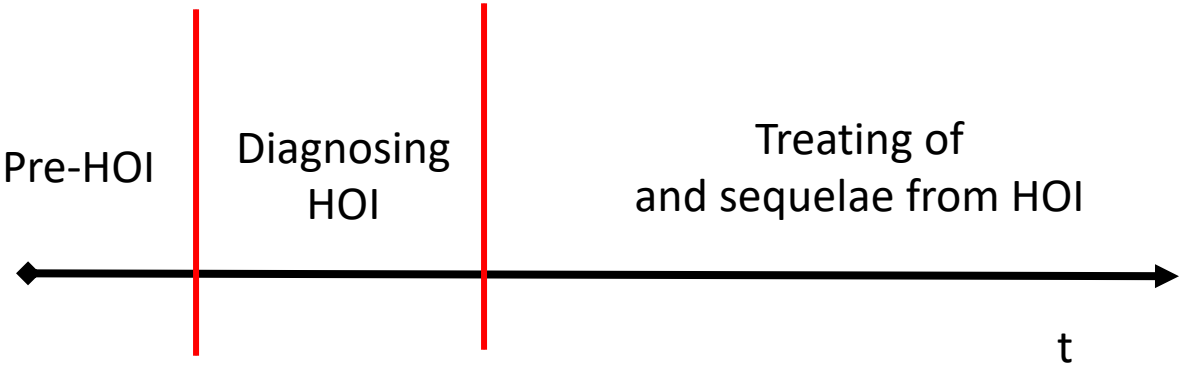
Randomly Selected Subject Population For Evaluation Cohort



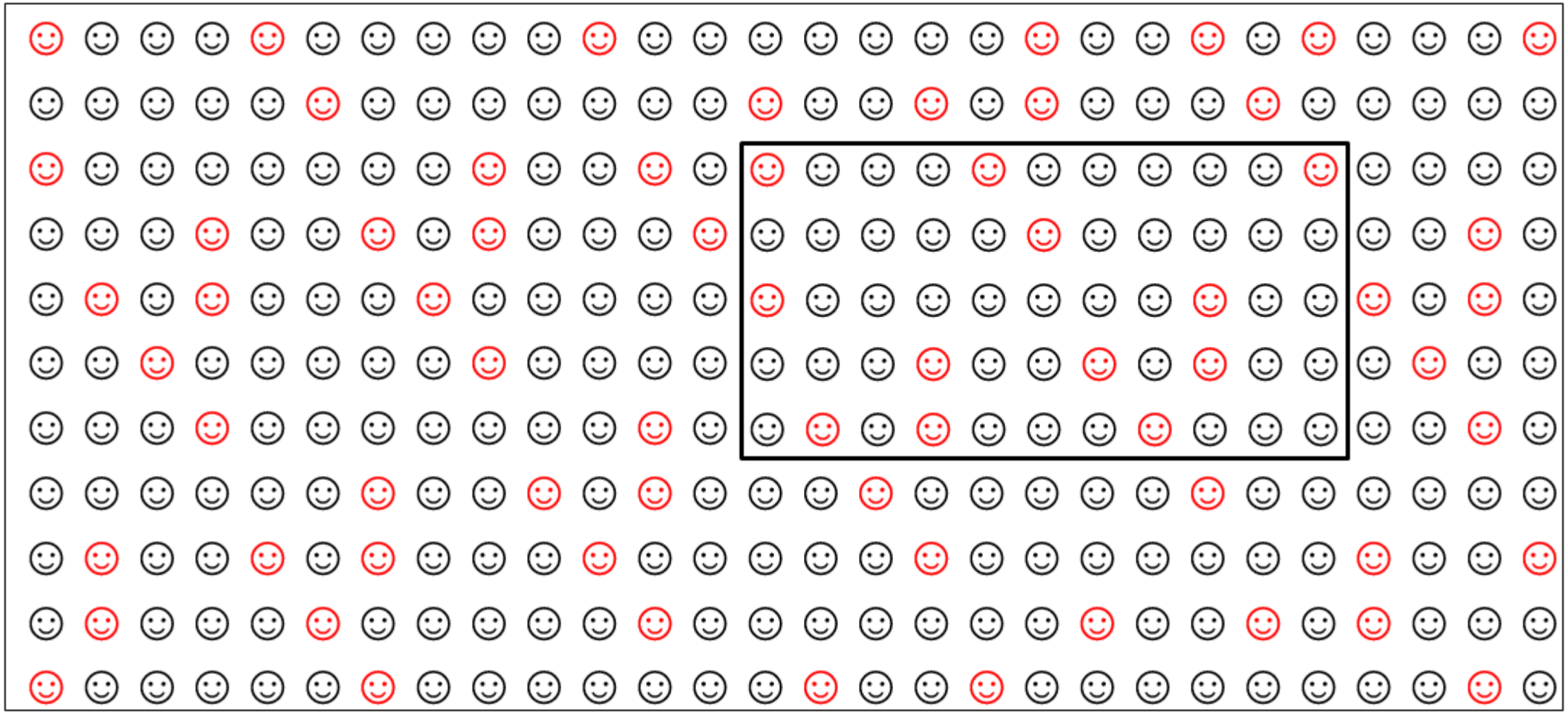
How the Model Works





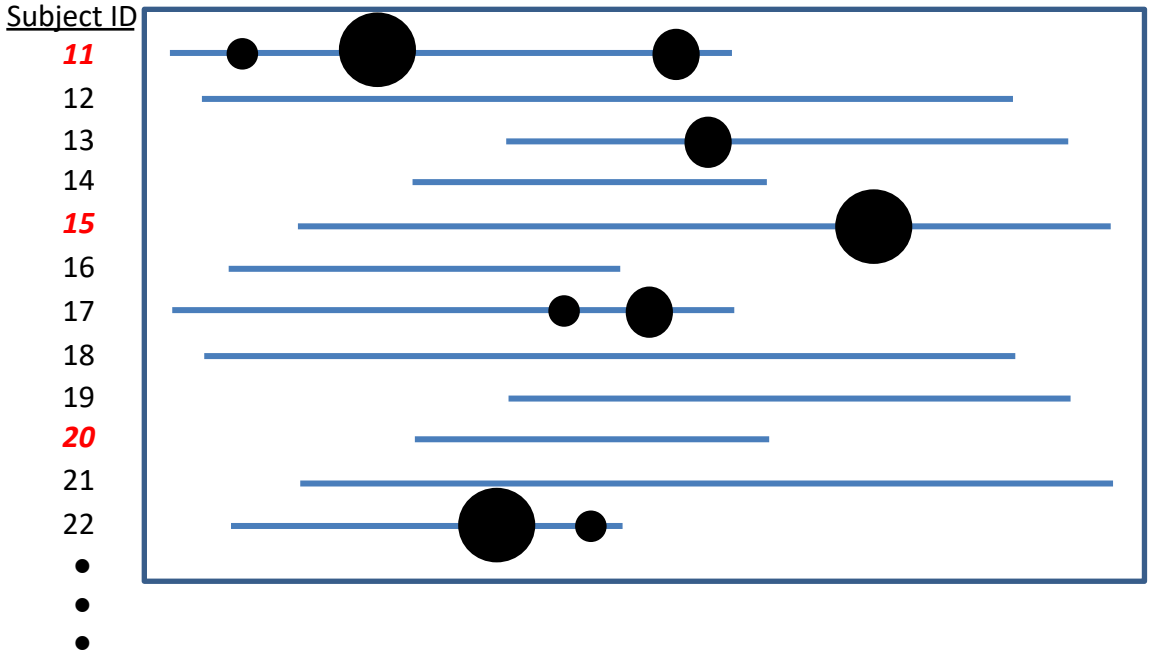


Testing Phenotype Algorithms



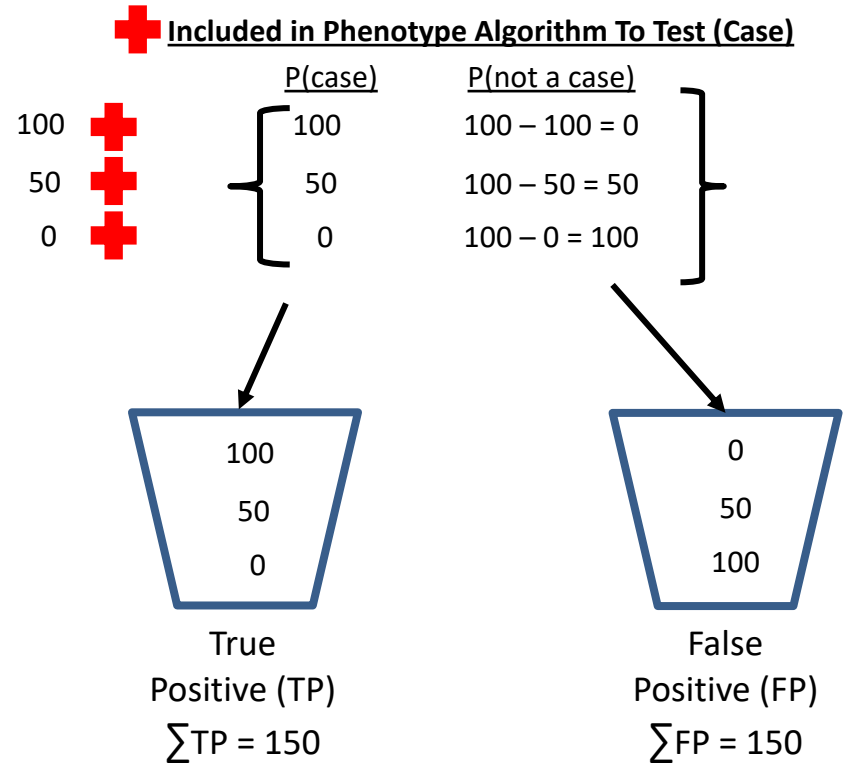
Assessing the Phenotype Algorithm, Part 1

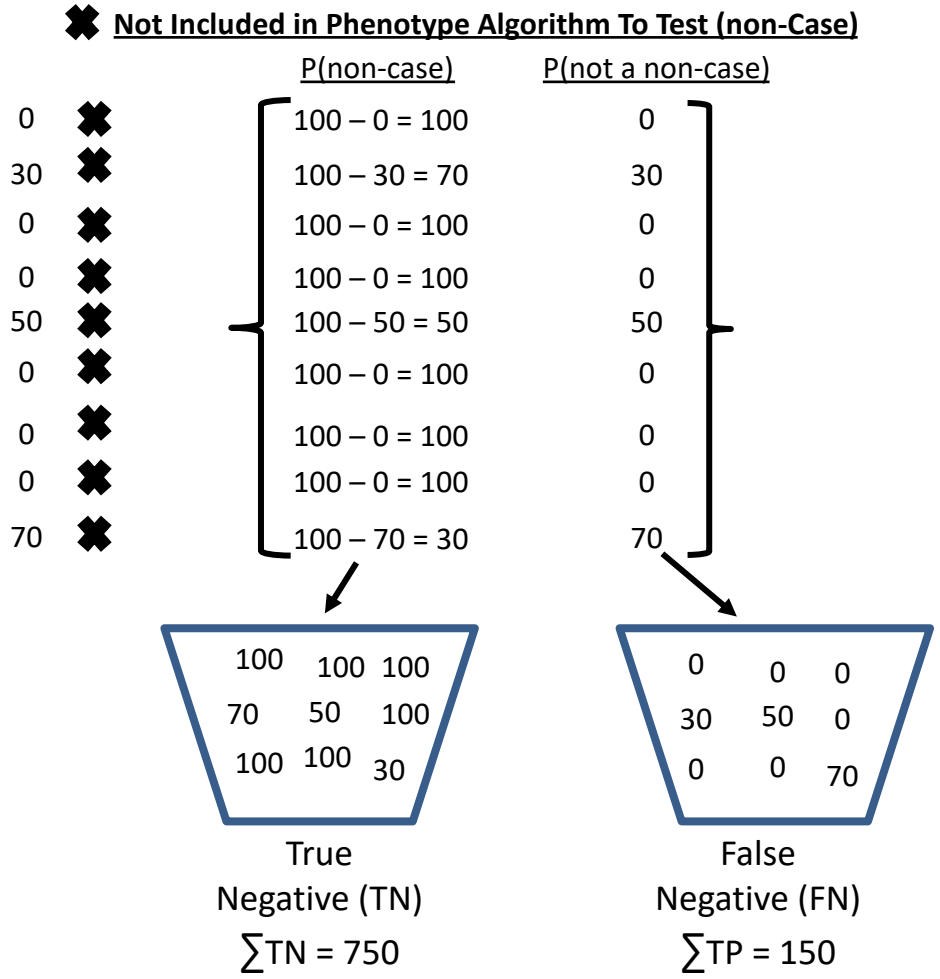
Randomly Selected Subject Population
For Evaluation Cohort with Predictors



Assessing the Phenotype Algorithm, Part 2

Subject ID	Predicted Probability of HOI	
11	100	✚
12	0	✖
13	30	✖
14	0	✖
15	50	✚
16	0	✖
17	50	✖
18	0	✖
19	0	✖
20	0	✚
21	0	✖
22	70	✖





Assessing the Phenotype Algorithm, Part 3

True Positive (TP)
∑TP = 150

False Positive (FP)
∑FP = 150

True Negative (TN)
∑TN = 750

False Negative (FN)
∑TP = 150

Confusion Matrix		Truth	
		Positive	Negative
Test	Positive	TP=150	FP=150
	Negative	FN=150	TN=750

Sensitivity = TP/(TP+FN) = 150/(150 + 150) = 0.50
Specificity = TN/(TN + FP) = 750/(750 + 150) = 0.83
Positive Predictive Value = TP/(TP + FP) = 150/(150 + 150) = 0.50
Negative Predictive Value = TN/(TN + FN) = 750/(750 + 150) = 0.83

Testing the Phenotypes

- Typical Phenotypes for MI:
- 1 X MI (Myocardial Infarction - SNOMED concept ID 22298006)
- 2 X MI, second MI diagnosis within 5 days of first MI diagnosis
- 1 X MI, In-patient
- 1 X MI, In-patient in first position

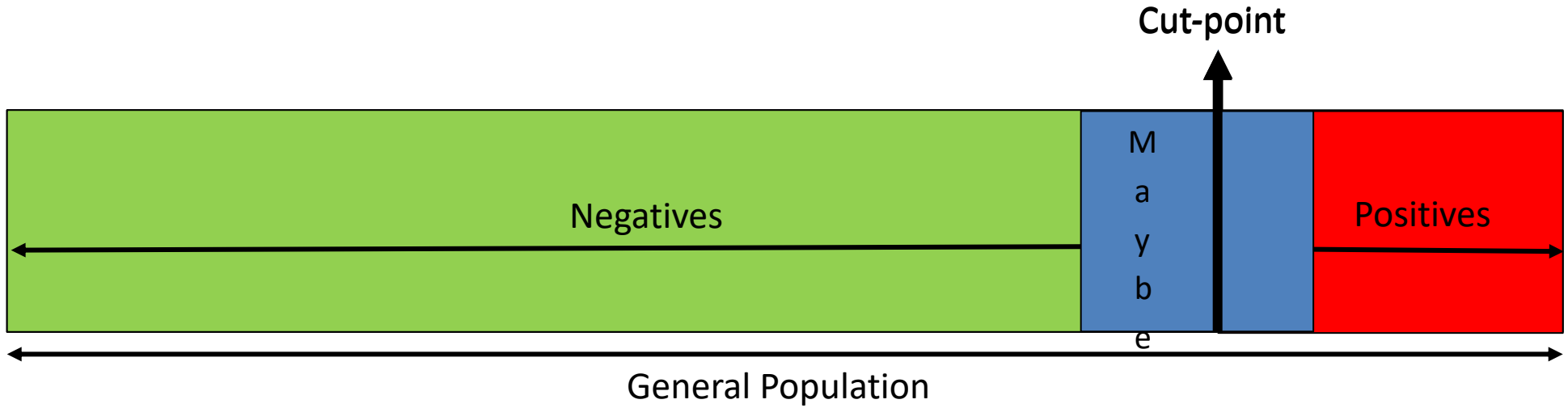
Comparisons with Published Validations

		PheValuator			Prior Studies				
Outcome	Algorithm	Sens	PPV	Spec	Sens	PPV	Spec	Author	N
Myocardial Infarction	>=1 x HOI, IP - 1st Position		94		NA	86	NA	Cutrona (2014)	153
	>=1 x HOI, IP - 1st Position		94		NA	93	NA	Choma (2009)	350
	>=1 x HOI, IP - 1st Position (MDCR)		91		NA	88	NA	Kiyota (2004)	2200
Ischemic Stroke	>=1 x HOI, IP - 1st Position (MDCR)	69	89	99	59	91	99	Kumamaru (2014)	15089
	>=1 x HOI, IP - 1st Position		87		NA	88	NA	Giroud (2015)	1680
	>=1 x HOI, IP - 1st Position		87		NA	88	NA	Lühdorf (2017)	3326
Diabetes	>= 1 x HOI	99	89	97	93	91	99	Crane (2006)	53
	>= 1 x HOI (MDCR)	99	91	96	79	71	94	Hebert (1999)	-
	>=1 x HOI, IP	42	92	99	67	83	99	So (2006)	93
Atrial Fibrillation	>= 1 x HOI		84		NA	78	NA	Go (2000)	50
	>=1 x HOI, IP	56	94	99	84	89	98	Alonso (2009)	125
	>=1 x HOI, IP		94		NA	91	NA	Antani (1996)	196

Sens - Sensitivity; PPV - Positive Predictive Value; Spec - Specificity; MDCR - Truven Medicare;
HOI - Health Outcome of Interest; IP - In-Patient; NA - Not analyzed

Limitations

- Sparse data for subjects
- Databases vary with overall level of detail
- Complex coding for conditions, e.g., MI v. T2DM



- Cutrona – 10% of patients with insufficient evidence
- Ryo – 7.5% of patients with insufficient evidence

Conclusions/Next Steps

- Using diagnostic predictive models to assess algorithm performance appears promising
- Having metrics for phenotype performance increases confidence in the use of observational data in research.
- Potential to use results of phenotype evaluation to correct/adjust our estimates

PheValuator

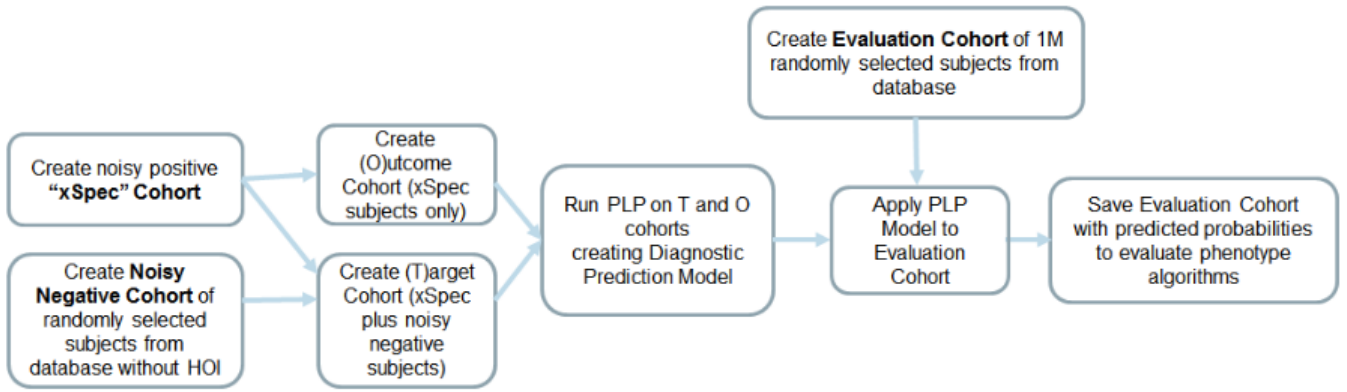
An R package for evaluating phenotype algorithms,

Introduction

The goal of PheValuator is to produce a large cohort of subjects each with a predicted probability for a specified health outcome of interest (HOI). This is achieved by developing a diagnostic predictive model for the HOI using the PatientLevelPrediction (PLP) R package and applying the model to a large, randomly selected population. These subjects can be used to test one or more phenotype algorithms.

Process Steps

The first step in the process, developing the evaluation cohort, is shown below:



https://github.com/OHDSI/phevaluator