Journey toward Patient-Level Prediction

Peter R. Rijnbeek, PhD

Department of Medical Informatics

Erasmus MC, Rotterdam, The Netherlands

Jenna Reps, PhD

Janssen Research and Development

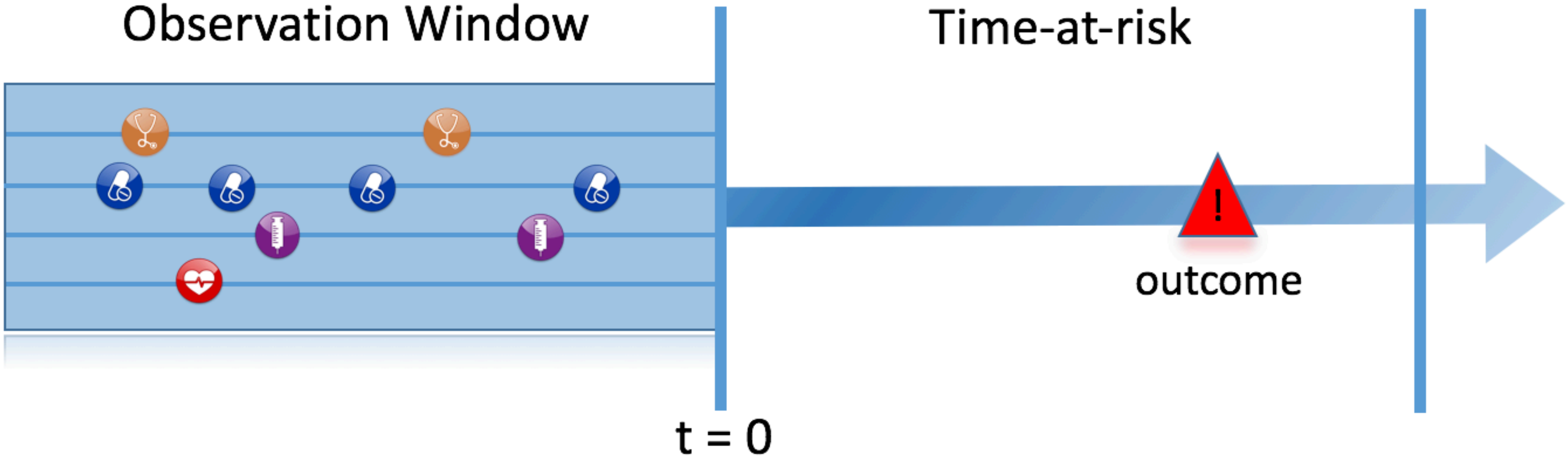
Peter R. Rijnbeek, PhD

Department of Medical Informatics

Erasmus MC, Rotterdam, The Netherlands

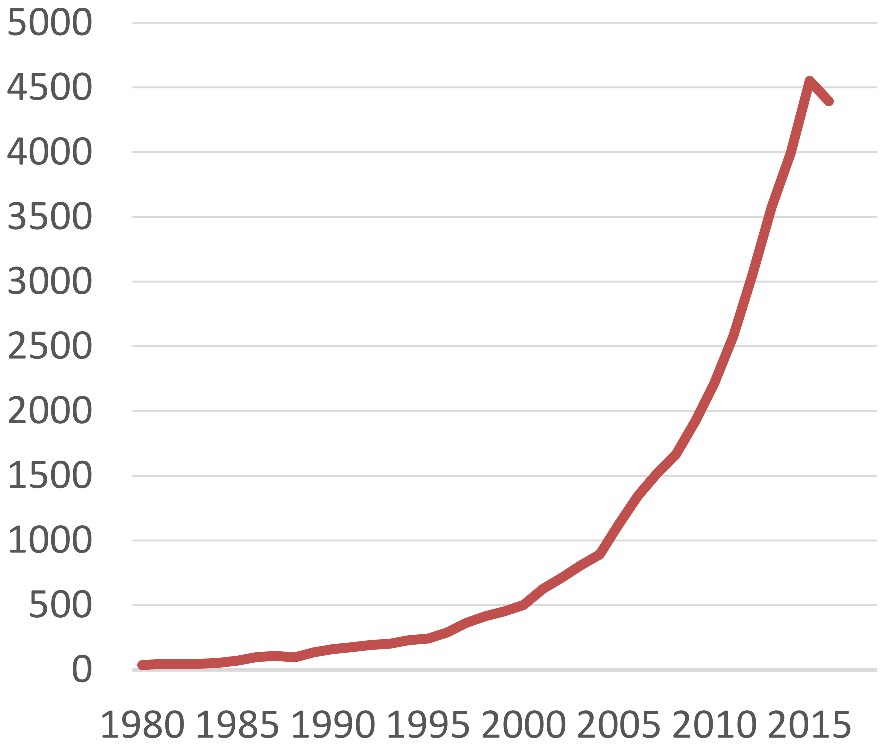
Prediction is difficult, especially about the future !
CERTIFICATE OF MEDICINE

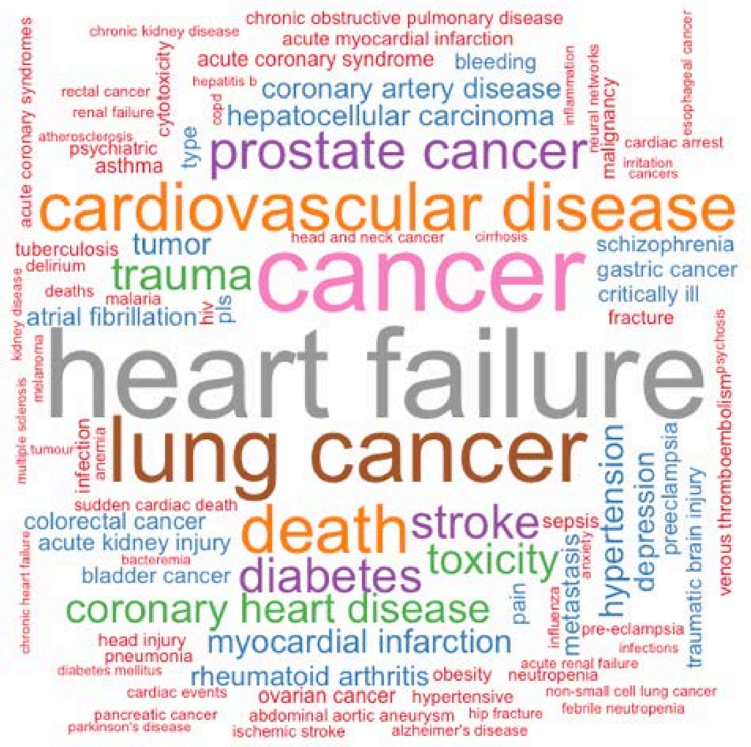
Problem definition



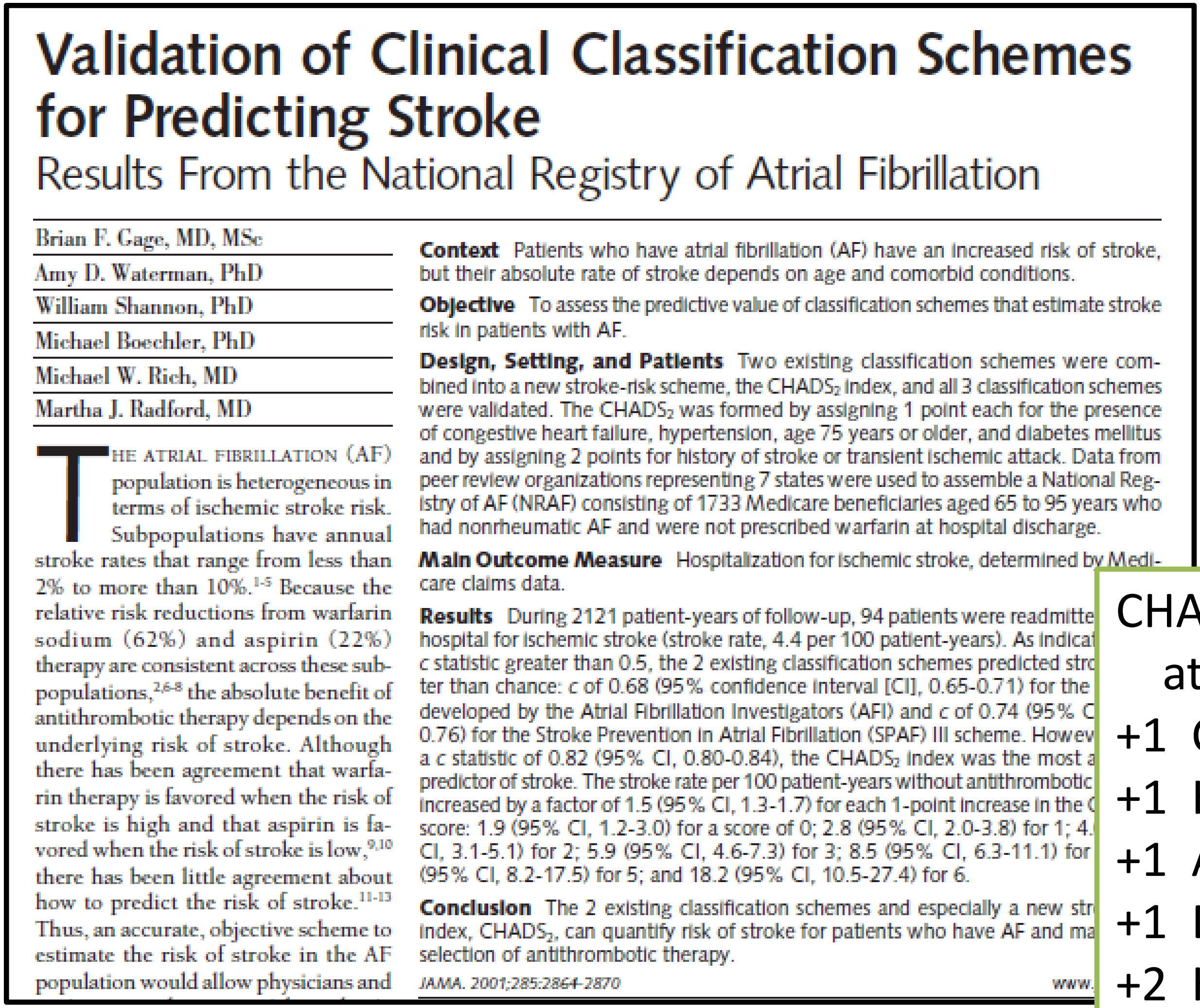
Among a population at risk (Depression), we aim to predict which patients at a defined moment in time (t=0) will experience some outcome (Stroke) during a time-at-risk (1 year). Prediction is done using only information about the patients in an observation window prior to that moment in time.

Growing interest in prediction modelling





Patient-level prediction models are already in clinical practice



Validation of Clinical Classification Schemes for Predicting Stroke

Results From the National Registry of Atrial Fibrillation

Brian F. Gage, MD, MSc
Amy D. Waterman, PhD
William Shannon, PhD
Michael Boechler, PhD
Michael W. Rich, MD
Martha J. Radford, MD

THE ATRIAL FIBRILLATION (AF) population is heterogeneous in terms of ischemic stroke risk. Subpopulations have annual stroke rates that range from less than 2% to more than 10%.[1-5] Because the relative risk reductions from warfarin sodium (62%) and aspirin (22%) therapy are consistent across these subpopulations,[2,6-8] the absolute benefit of antithrombotic therapy depends on the underlying risk of stroke. Although there has been agreement that warfarin therapy is favored when the risk of stroke is high and that aspirin is favored when the risk of stroke is low,[9,10] there has been little agreement about how to predict the risk of stroke.[11-13] Thus, an accurate, objective scheme to estimate the risk of stroke in the AF population would allow physicians and

Context Patients who have atrial fibrillation (AF) have an increased risk of stroke, but their absolute rate of stroke depends on age and comorbid conditions.

Objective To assess the predictive value of classification schemes that estimate stroke risk in patients with AF.

Design, Setting, and Patients Two existing classification schemes were combined into a new stroke-risk scheme, the $CHADS_2$ index, and all 3 classification schemes were validated. The $CHADS_2$ was formed by assigning 1 point each for the presence of congestive heart failure, hypertension, age 75 years or older, and diabetes mellitus and by assigning 2 points for history of stroke or transient ischemic attack. Data from peer review organizations representing 7 states were used to assemble a National Registry of AF (NRAF) consisting of 1733 Medicare beneficiaries aged 65 to 95 years who had nonrheumatic AF and were not prescribed warfarin at hospital discharge.

Main Outcome Measure Hospitalization for ischemic stroke, determined by Medicare claims data.

Results During 2121 patient-years of follow-up, 94 patients were readmitte
hospital for ischemic stroke (stroke rate, 4.4 per 100 patient-years). As indicat
c statistic greater than 0.5, the 2 existing classification schemes predicted stro
ter than chance: *c* of 0.68 (95% confidence interval [CI], 0.65-0.71) for the
developed by the Atrial Fibrillation Investigators (AFI) and *c* of 0.74 (95% C
0.76) for the Stroke Prevention in Atrial Fibrillation (SPAF) III scheme. Howev
a *c* statistic of 0.82 (95% CI, 0.80-0.84), the $CHADS_2$ index was the most a
predictor of stroke. The stroke rate per 100 patient-years without antithrombotic
increased by a factor of 1.5 (95% CI, 1.3-1.7) for each 1-point increase in the C
score: 1.9 (95% CI, 1.2-3.0) for a score of 0; 2.8 (95% CI, 2.0-3.8) for 1; 4.
CI, 3.1-5.1) for 2; 5.9 (95% CI, 4.6-7.3) for 3; 8.5 (95% CI, 6.3-11.1) for
(95% CI, 8.2-17.5) for 5; and 18.2 (95% CI, 10.5-27.4) for 6.

Conclusion The 2 existing classification schemes and especially a new str
index, $CHADS_2$, can quantify risk of stroke for patients who have AF and ma
selection of antithrombotic therapy.

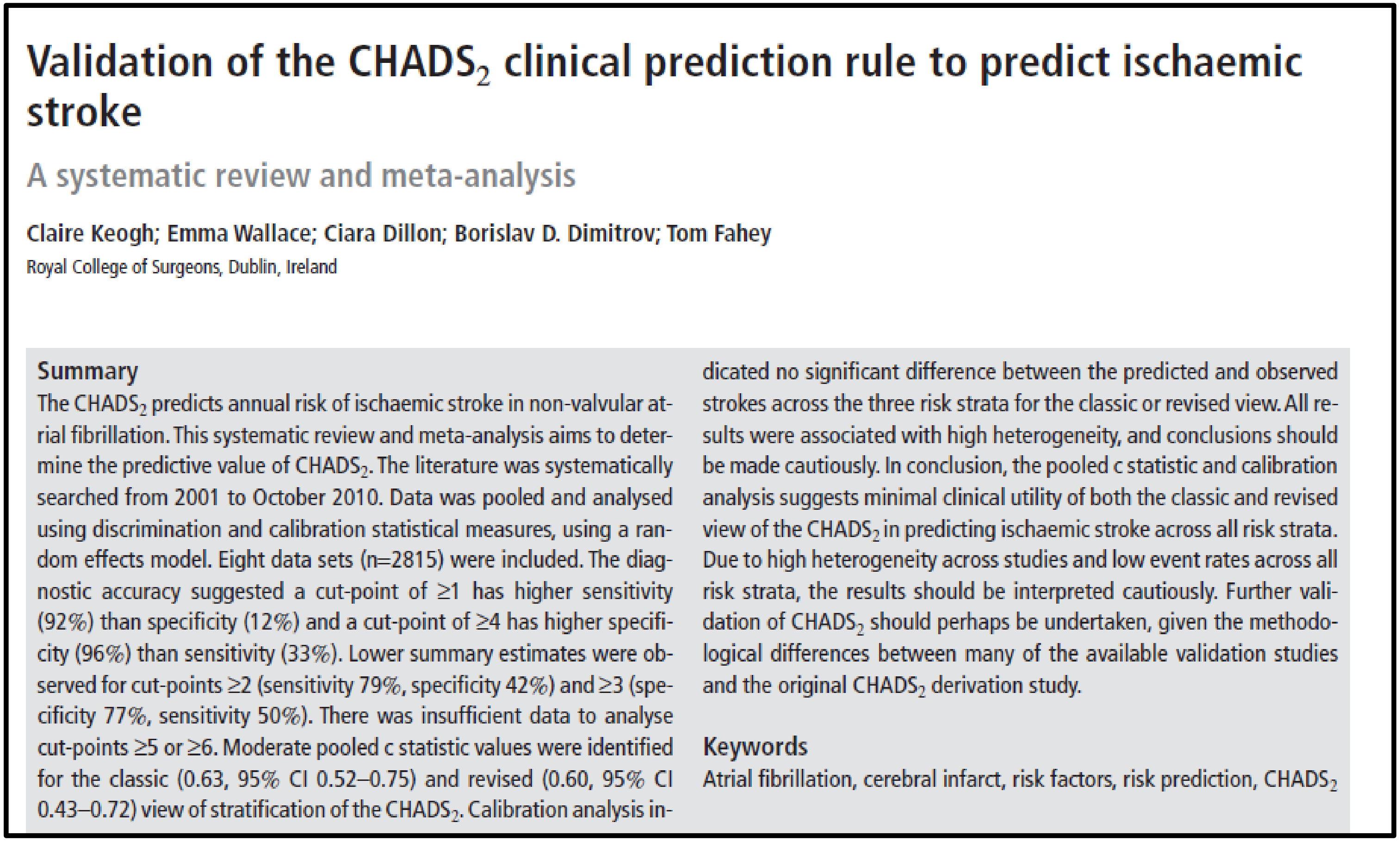
JAMA. 2001;285:2864-2870 www.

CHADS2 for patients with atrial fibrillation:

- +1 Congestive heart failure
- +1 Hypertension
- +1 Age >= 75
- +1 Diabetes mellitus
- +2 History of transient ischemic attack

JAMA, 2001; 285: 2864-2870

Evaluating the predictive accuracy of CHADS2



Validation of the $CHADS_2$ clinical prediction rule to predict ischaemic stroke

A systematic review and meta-analysis

Claire Keogh; Emma Wallace; Ciara Dillon; Borislav D. Dimitrov; Tom Fahey
Royal College of Surgeons, Dublin, Ireland

Summary

The $CHADS_2$ predicts annual risk of ischaemic stroke in non-valvular atrial fibrillation. This systematic review and meta-analysis aims to determine the predictive value of $CHADS_2$. The literature was systematically searched from 2001 to October 2010. Data was pooled and analysed using discrimination and calibration statistical measures, using a random effects model. Eight data sets (n=2815) were included. The diagnostic accuracy suggested a cut-point of ≥1 has higher sensitivity (92%) than specificity (12%) and a cut-point of ≥4 has higher specificity (96%) than sensitivity (33%). Lower summary estimates were observed for cut-points ≥2 (sensitivity 79%, specificity 42%) and ≥3 (specificity 77%, sensitivity 50%). There was insufficient data to analyse cut-points ≥5 or ≥6. Moderate pooled c statistic values were identified for the classic (0.63, 95% CI 0.52–0.75) and revised (0.60, 95% CI 0.43–0.72) view of stratification of the $CHADS_2$. Calibration analysis indicated no significant difference between the predicted and observed strokes across the three risk strata for the classic or revised view. All results were associated with high heterogeneity, and conclusions should be made cautiously. In conclusion, the pooled c statistic and calibration analysis suggests minimal clinical utility of both the classic and revised view of the $CHADS_2$ in predicting ischaemic stroke across all risk strata. Due to high heterogeneity across studies and low event rates across all risk strata, the results should be interpreted cautiously. Further validation of $CHADS_2$ should perhaps be undertaken, given the methodological differences between many of the available validation studies and the original $CHADS_2$ derivation study.

Keywords

Atrial fibrillation, cerebral infarct, risk factors, risk prediction, $CHADS_2$

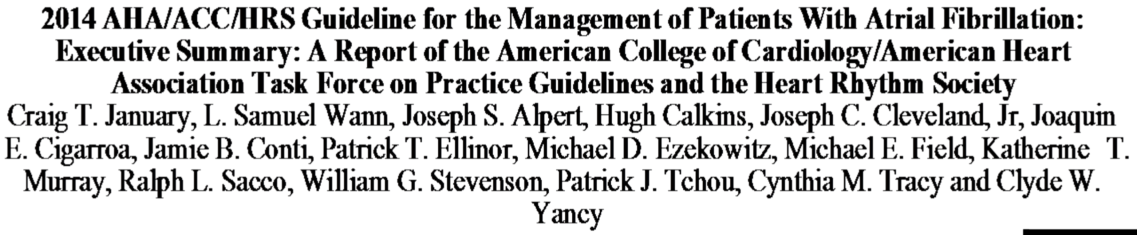
Thromb Haemost 2011; 106: 528–538

Current Stroke Guidelines

Circulation

JOURNAL OF THE AMERICAN HEART ASSOCIATION

American Heart Association



2014 AHA/ACC/HRS Guideline for the Management of Patients With Atrial Fibrillation: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society

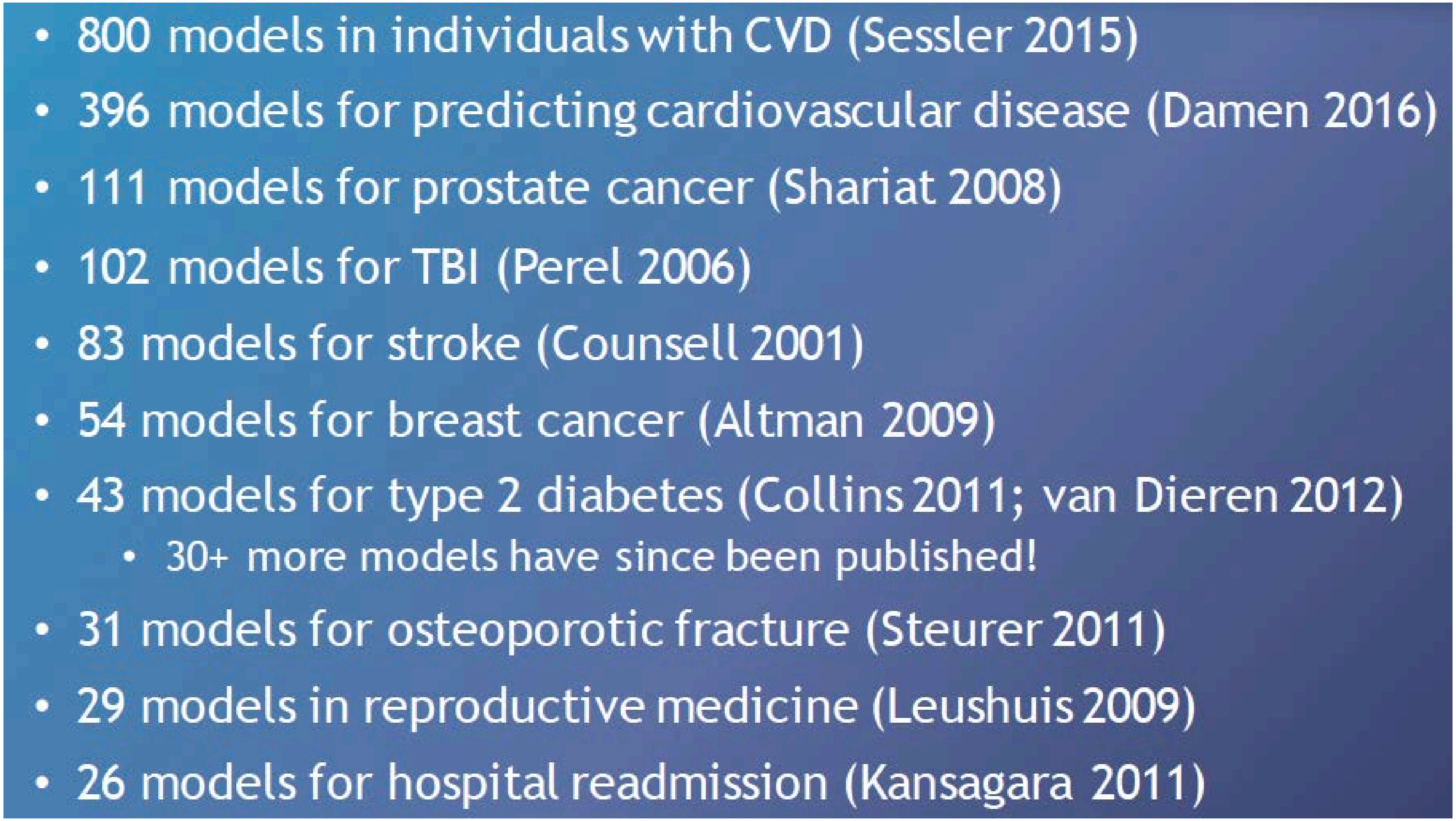
Craig T. January, L. Samuel Wann, Joseph S. Alpert, Hugh Calkins, Joseph C. Cleveland, Jr, Joaquin E. Cigarroa, Jamie B. Conti, Patrick T. Ellinor, Michael D. Ezekowitz, Michael E. Field, Katherine T. Murray, Ralph L. Sacco, William G. Stevenson, Patrick J. Tchou, Cynthia M. Tracy and Clyde W. Yancy

Recommendation:

In patients with **nonvalvular atrial fibrillation**, the CHA_2DS_2-VASc score is recommended for assessment of stroke risk

CHA_2DS_2-VASc Risk	Score
CHF or LVEF $\leq$ 40%	1
Hypertension	1
Age $\geq$ 75	2
Diabetes	1
Stroke/TIA/ Thromboembolism	2
Vascular Disease	1
Age 65 - 74	1
Female	1

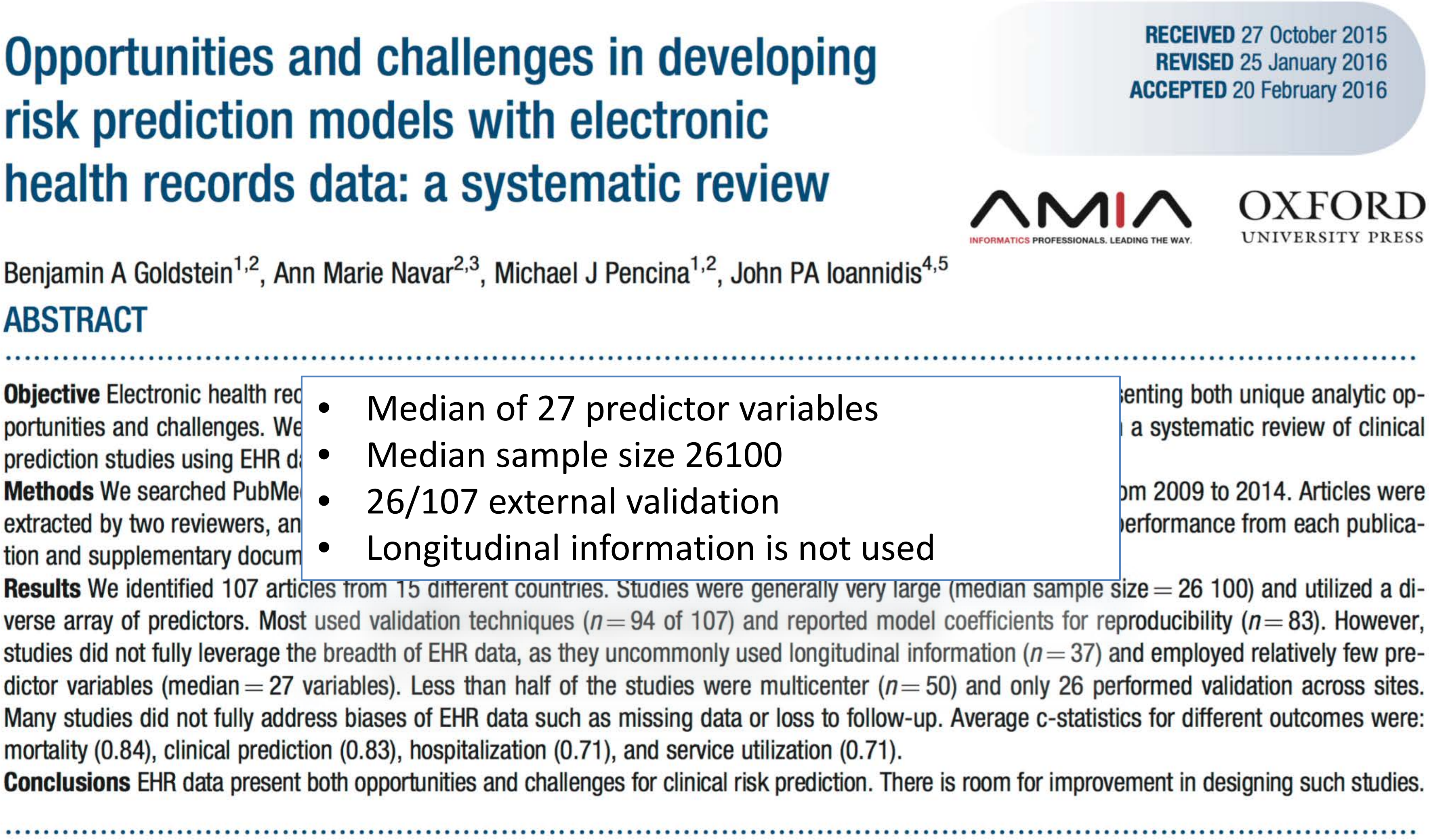
Reviews of published prediction models



- 800 models in individuals with CVD (Sessler 2015)
- 396 models for predicting cardiovascular disease (Damen 2016)
- 111 models for prostate cancer (Shariat 2008)
- 102 models for TBI (Perel 2006)
- 83 models for stroke (Counsell 2001)
- 54 models for breast cancer (Altman 2009)
- 43 models for type 2 diabetes (Collins 2011; van Dieren 2012)
  - 30+ more models have since been published!
- 31 models for osteoporotic fracture (Steurer 2011)
- 29 models in reproductive medicine (Leushuis 2009)
- 26 models for hospital readmission (Kansagara 2011)

Courtesy of Gary Collins

Current status of prediction modelling



Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review

RECEIVED 27 October 2015
REVISED 25 January 2016
ACCEPTED 20 February 2016

AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.
OXFORD
UNIVERSITY PRESS

Benjamin A Goldstein[1,2], Ann Marie Navar[2,3], Michael J Pencina[1,2], John PA Ioannidis[4,5]

ABSTRACT

Objective Electronic health rec
senting both unique analytic opportunities and challenges. We
a systematic review of clinical prediction studies using EHR d

Methods We searched PubMe
om 2009 to 2014. Articles were extracted by two reviewers, an
erformance from each publication and supplementary docum

Results We identified 107 articles from 15 different countries. Studies were generally very large (median sample size = 26 100) and utilized a diverse array of predictors. Most used validation techniques ($n = 94$ of 107) and reported model coefficients for reproducibility ($n = 83$). However, studies did not fully leverage the breadth of EHR data, as they uncommonly used longitudinal information ($n = 37$) and employed relatively few predictor variables (median = 27 variables). Less than half of the studies were multicenter ($n = 50$) and only 26 performed validation across sites. Many studies did not fully address biases of EHR data such as missing data or loss to follow-up. Average c-statistics for different outcomes were: mortality (0.84), clinical prediction (0.83), hospitalization (0.71), and service utilization (0.71).

Conclusions EHR data present both opportunities and challenges for clinical risk prediction. There is room for improvement in designing such studies.

- Median of 27 predictor variables
- Median sample size 26100
- 26/107 external validation
- Longitudinal information is not used

Goldstein BA, J Am Med Inform Assoc. 2016.

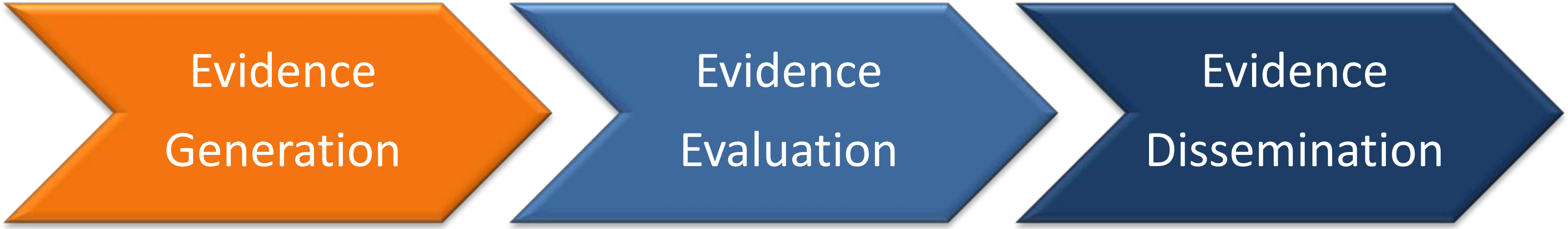
Current status of prediction modelling

- Inadequate internal validation
- Small sets of features
- Incomplete dissemination of model and results
- No transportability assessment
- Impact on clinical decision making unknown

→ Relatively few prediction models are used in clinical practice

Mission for Patient-Level Prediction

OHDSI aims to develop a systematic process to learn and evaluate large-scale patient-level prediction models using observational health data in a data network

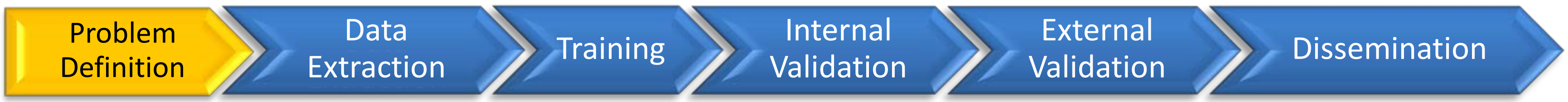


Evidence Generation

Evidence Evaluation

Evidence Dissemination

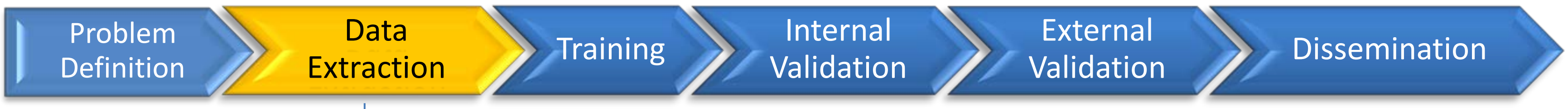
Prediction Model Development

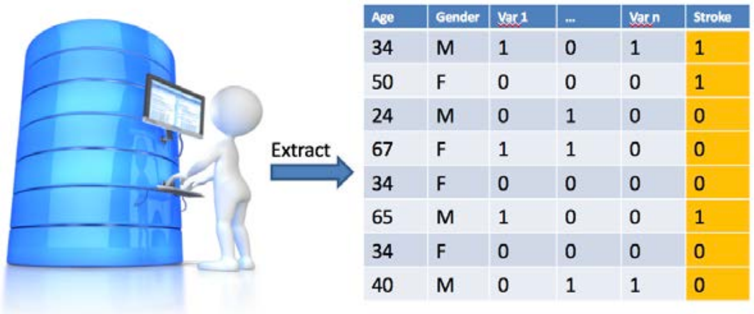


Problem pre-specification. A study protocol should unambiguously pre-specify the planned analyses.

Transparency. Others should be able to reproduce a study in every detail using the provided information. All analysis code should be made available as open source on the OHDSI Github.

Prediction Model Development





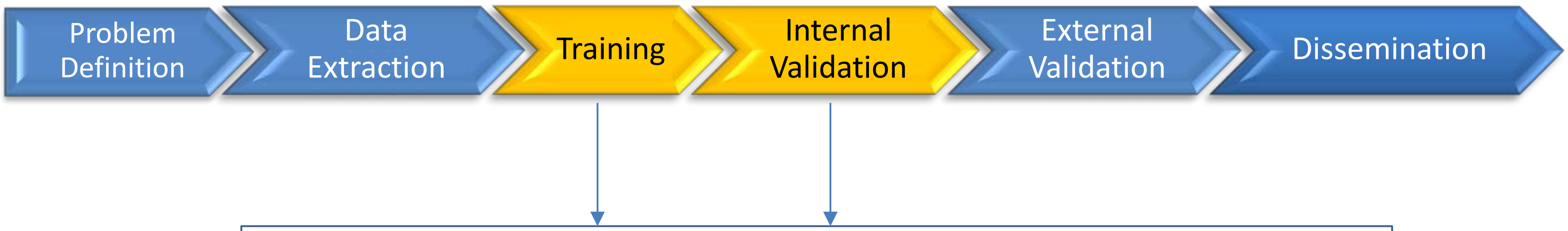
Age	Gender	Var 1	...	Var n	Stroke
34	M	1	0	1	1
50	F	0	0	0	1
24	M	0	1	0	0
67	F	1	1	0	0
34	F	0	0	0	0
65	M	1	0	0	1
34	F	0	0	0	0
40	M	0	1	1	0

Data is extracted from the OMOP CDM using the Feature Extraction R-Package.

Data characterization is required before modelling. Tools are being developed in the community to facilitate this.

A data cleaning step is recommended, e.g. to remove outliers in lab values.

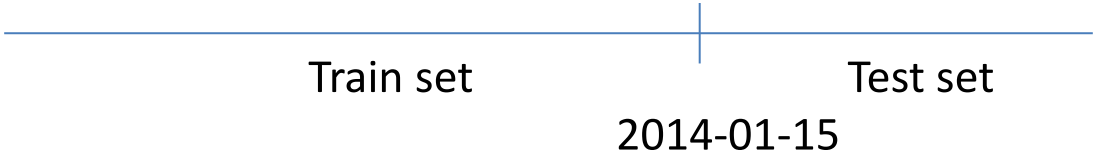
Prediction Model Development



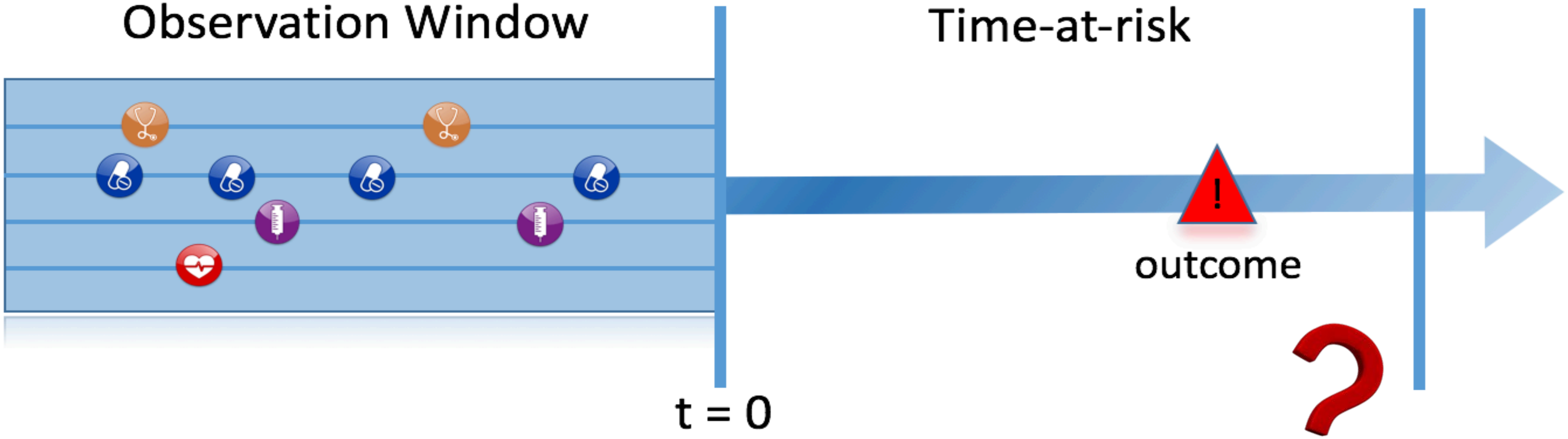
Model training and **Internal validation** is done using a train test split:

1. Person split: examples are assigned randomly to the train or test set, or

2. Time split: a split is made at a moment in time (temporal validation)



Model Training



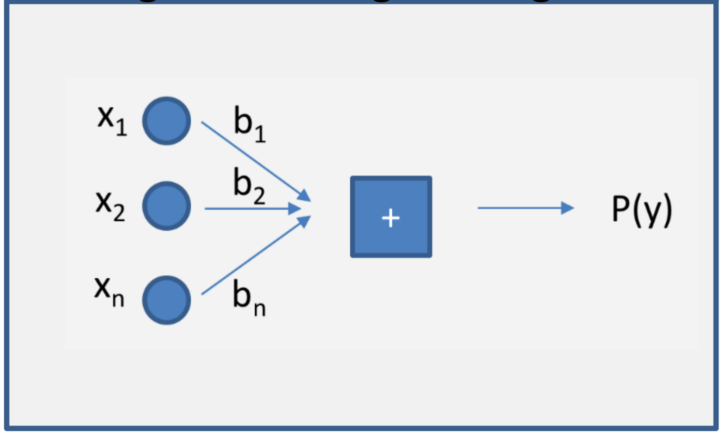
1. Which models?

2. How to evaluate the model?

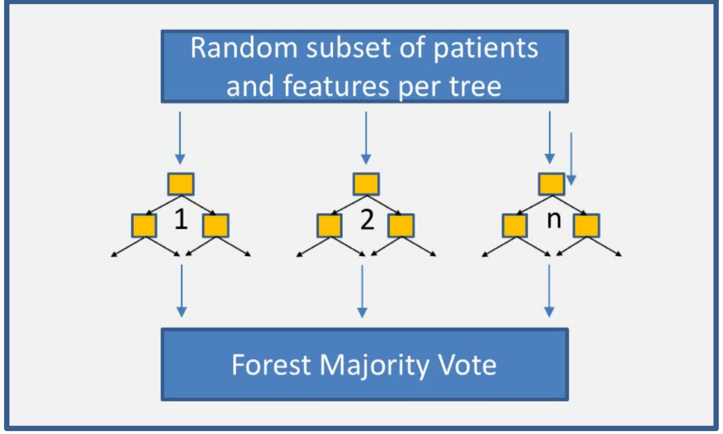
Models

Model training is an empirical process in which multiple models are compared

Regularized Logistic Regression

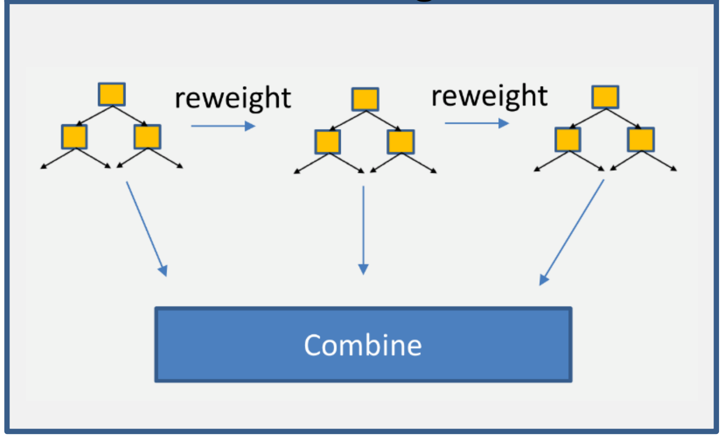


Random Forest


Random subset of patients and features per tree

Forest Majority Vote

Gradient Boosting Machines

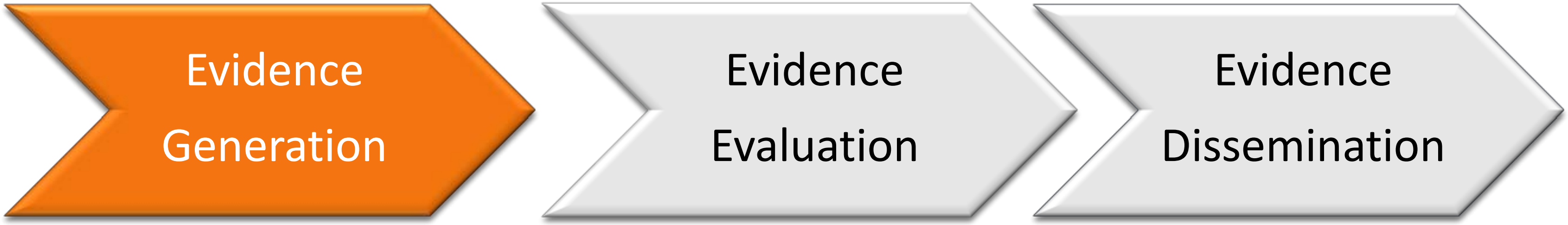


Combine

Many other models for example:

K-nearest neighbors
Naïve Bayes
Support Vector Machines
Etc.

Patient-Level Prediction Roadmap



Evidence Generation

Evidence Evaluation

Evidence Dissemination

Protocol Sharing
CDM Extractions
Code Sharing
Train / Test split

Model Validation

What makes a good model?

Discrimination: differentiates between those with and without the event, i.e. predicts higher probabilities for those with the event compared to those who don’t experience the event

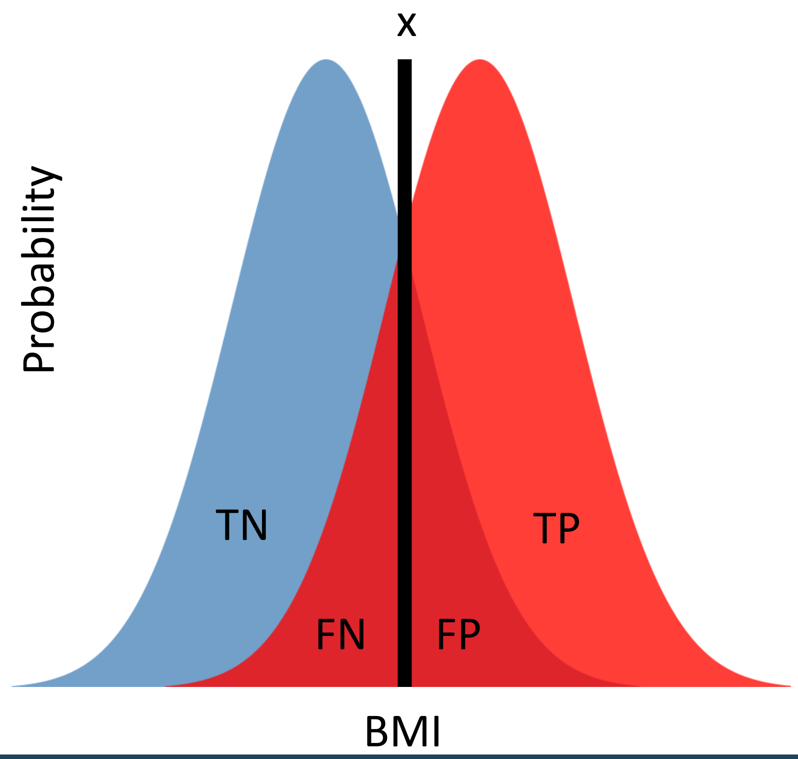
Calibration: estimated probabilities are close to the observed frequency

How to assess discrimination?

Suppose our classifier is simply BMI > x.

Both classes (blue = 0 , red = 1) have their own probability distribution of BMI

The choice of X then determines how sensitive or specific our algorithm is.



		Predicted	
		1	0
Observed	1		
	0		

True Positive Rate (TPR) = TP / (TP + FN)

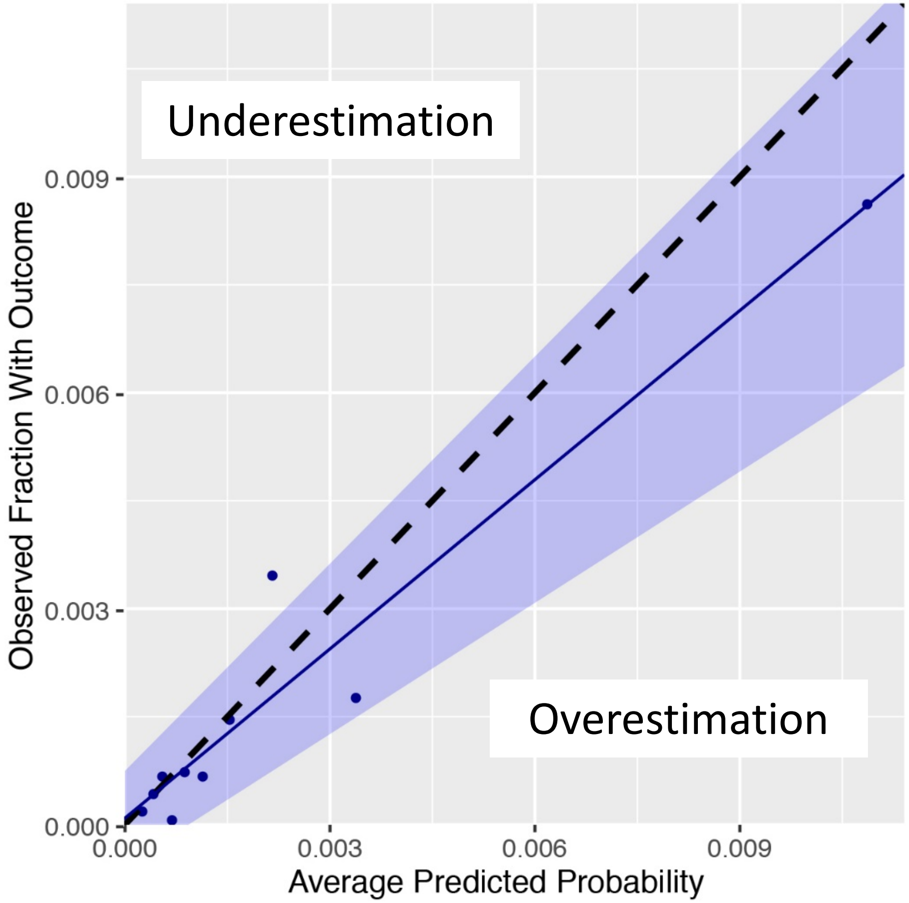
False Positive Rate (FPR) = FP / (FP + TN)

Receiver Operator Curve (ROC)

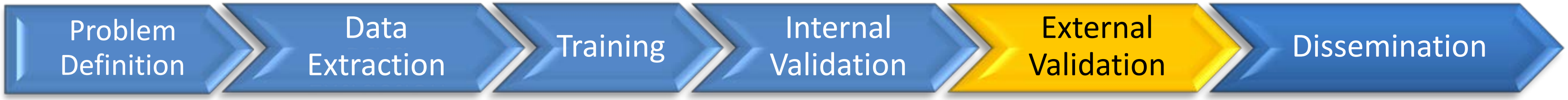
The Receiver Operator Curve (ROC) is developed during World War II for the analysis of radar images. Radar operators had to decide whether a blip on the screen represented an enemy target, a friendly ship, or just noise.

Calibration Assessment

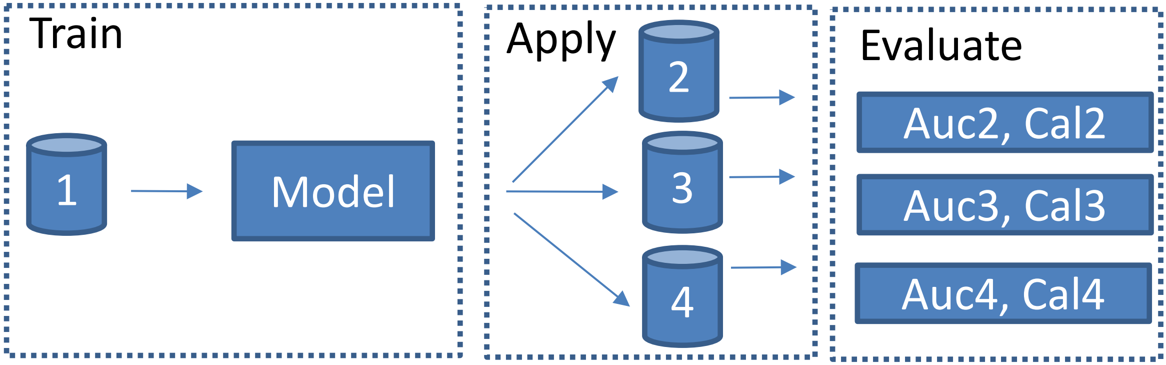
How close is the average predicted probability to the observed fraction with the outcome?



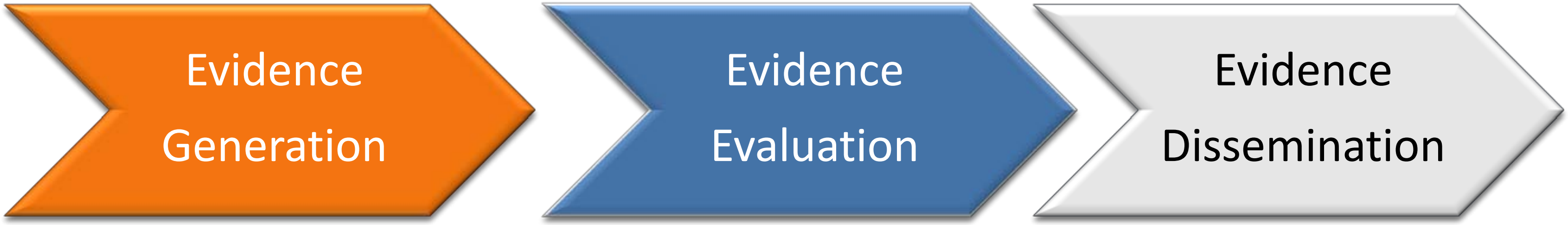
External Validation



External validation is performed using data from multiple populations not used for training.



Patient-Level Prediction Roadmap



Evidence Generation

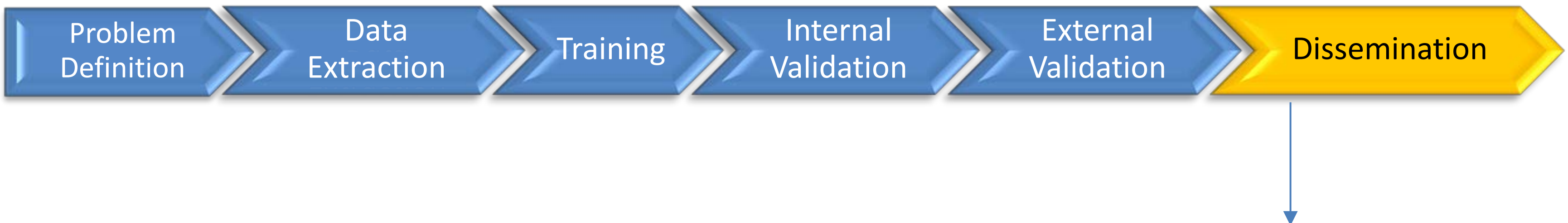
Evidence Evaluation

Evidence Dissemination

Protocol Sharing
CDM Extractions
Code Sharing
Train / Test split

Standardized Process
Discrimination
Calibration
External Validation

Dissemination



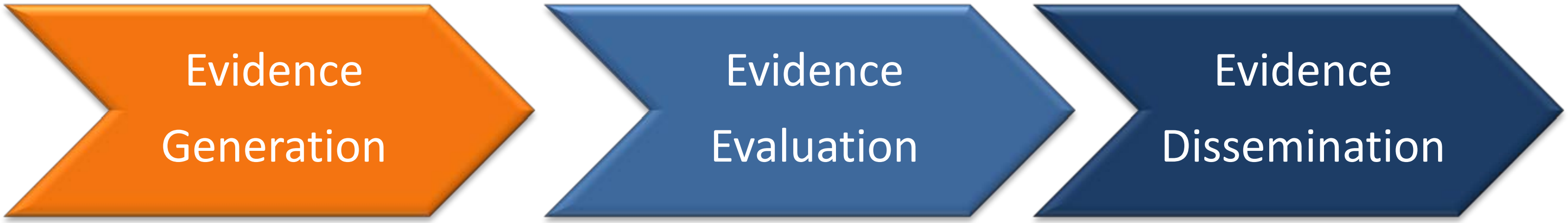
Dissemination of study results should follow the minimum requirements as stated in the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement [1].

- Internal and external validation
- Sharing of full model details
- Sharing of all analyses code to allow full reproducibility

→ Website to share protocol, code, models and results for all databases

[1] Moons, KG et al. Ann Intern Med. 2015;162(1):W1-73

Patient-Level Prediction Roadmap



Evidence Generation

Evidence Evaluation

Evidence Dissemination

Protocol Sharing
CDM Extractions
Code Sharing
Train / Test split

Standardization
Discrimination
Calibration
External Validation

Publications (TRIPOD)
Model sharing
Full transparency

OHDSI

Observational Health Data Sciences and Informatics

Large-scale patient-level prediction

A case study: prediction in patients with Pharmaceutically Treated Depression

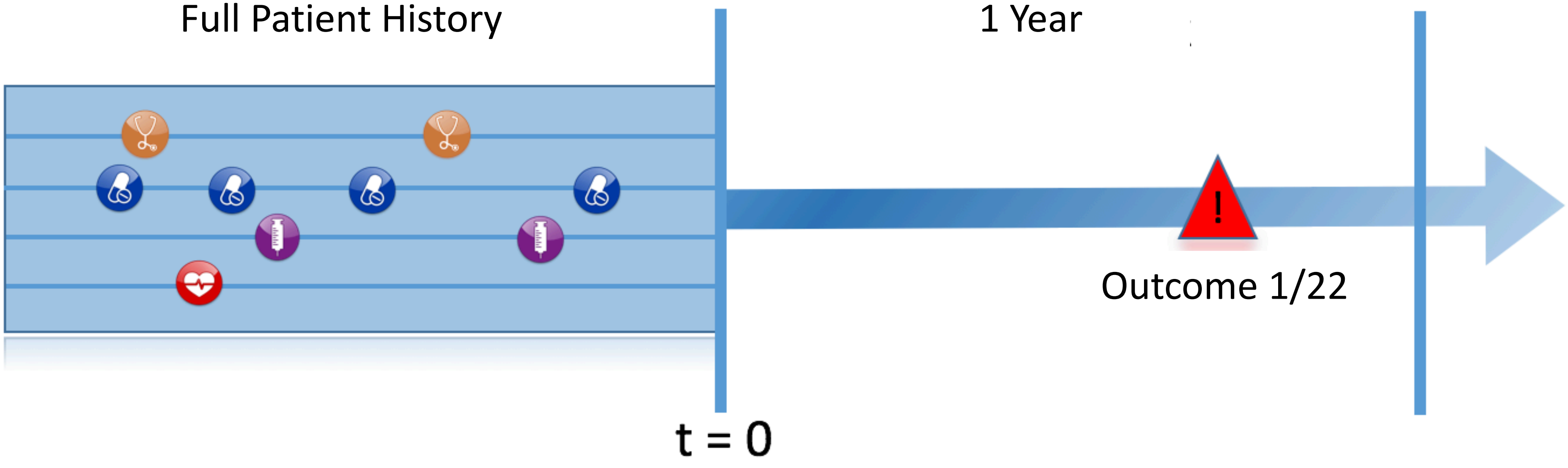
Objectives

- Assess the feasibility of large-scale predictive model development

- Investigate the performance of different classifiers across the outcomes and databases

- Initiate an assessment across the OHDSI data network

Problem definition



First Pharmaceutically Treated Depression

Among patients **in 4 different databases**, we aim to develop prediction models to predict which patients at a defined moment in time (**First Pharmaceutically Treated Depression Event**) will experience one out of **22 different outcomes** during a time-at-risk (**1 year**). Prediction is done using **all demographics, conditions, and drug use** data prior to that moment in time.

At Risk Cohort Definition

Patients are included in the cohort of interest at the date of the first occurrence of Pharmaceutically Treated Depression if the following inclusion criteria apply:

1. At least 365 days of history

2. At least 365 days of follow-up or the occurrence of the outcome of interest

3. No occurrence of the event prior to the index date

Setting

Databases

Database	Depression	Stroke
CCAE	659402	1351
MDCD	79818	356
MDCR	57839	874
OPTUM	363051	1183

Data extraction

- All demographics, conditions, drugs
- All 22 outcome cohorts

Training and testing

- Time split for training and testing
- Transportability for Stroke

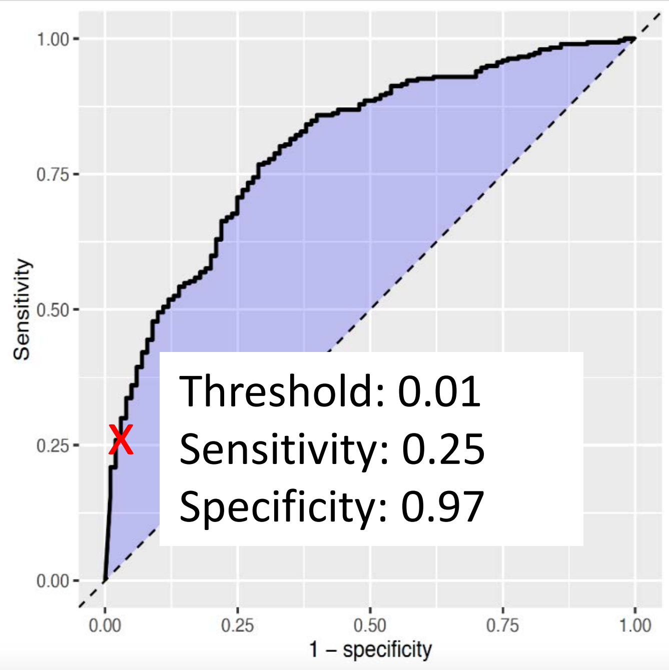
Models

- Gradient Boosting
- Random Forest
- Regularized Regression

Outcomes
Acute liver injury
Acute myocardial infarction
Alopecia
Constipation
Decreased libido
Delirium
Diarrhea
Fracture
Gastrointestinal hemhorrage
Hyperprolactinemia
Hyponatremia
Hypotension
Hypothyroidism
Insomnia
Nausea
Open-angle glaucoma
Seizure
Stroke
Suicide and suicidal ideation
Tinnitus
Ventricular arrhythmia and sudden cardiac death
Vertigo

Regularized Regression on CCAE

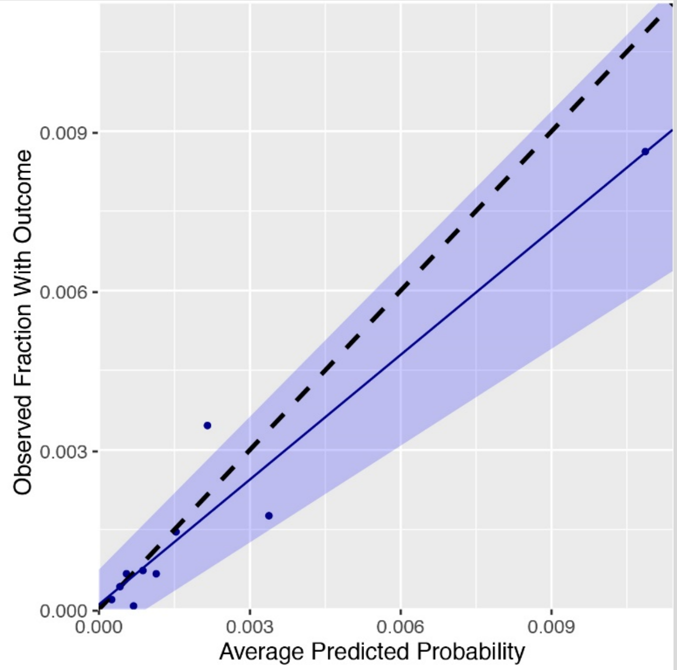
Receiver Operator Curve



Threshold: 0.01
Sensitivity: 0.25
Specificity: 0.97

AUC = 0.797

Calibration plot



Slope = 0.783

So what IS the model?

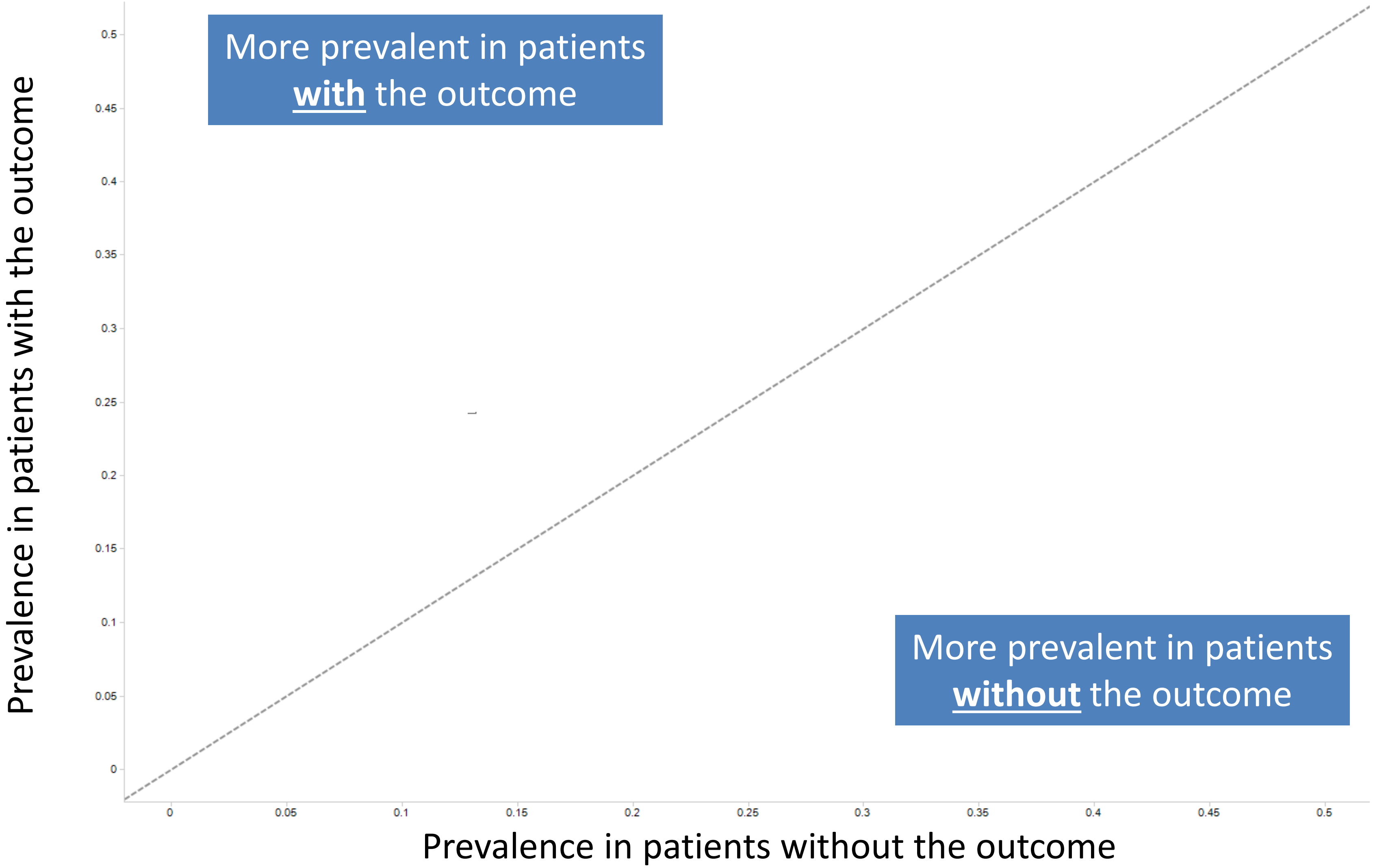
Reminder:

CHA_2DS_2-VASc is a model in clinical practices, but it was designed and tested for patients with Atrial Fibrillation to predict stroke, not for patients with depression and not for incident strokes….

The variables in this score were:
Age, Gender, Congestive Heart Failure, Hypertension, Diabetes, Vascular disease

→ Did our model pick those variables automatically from the data?

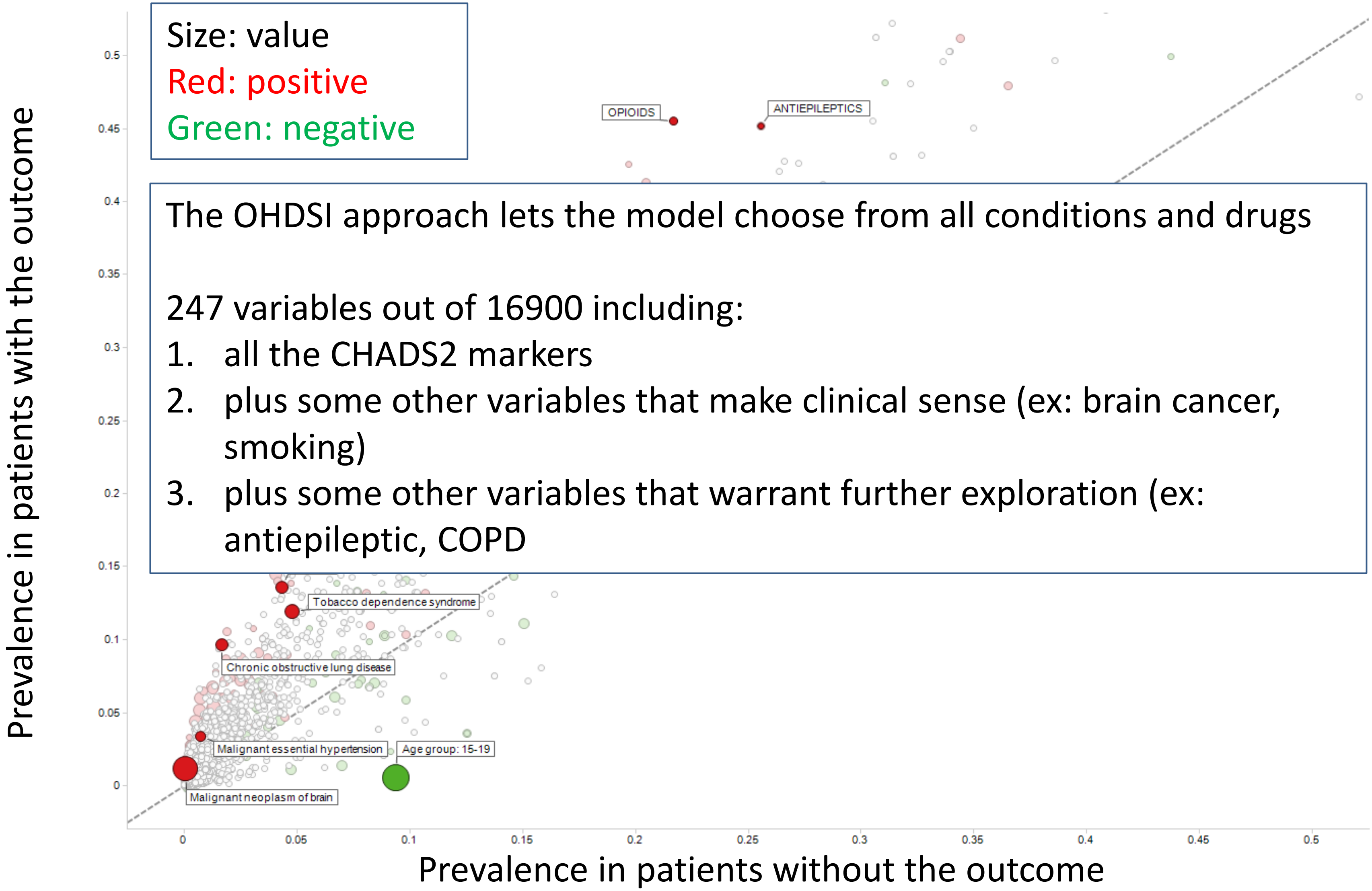
CHA$_2$DS$_2$-VASc variables



More prevalent in patients **with** the outcome

More prevalent in patients **without** the outcome

All variables explored in a large-scale model



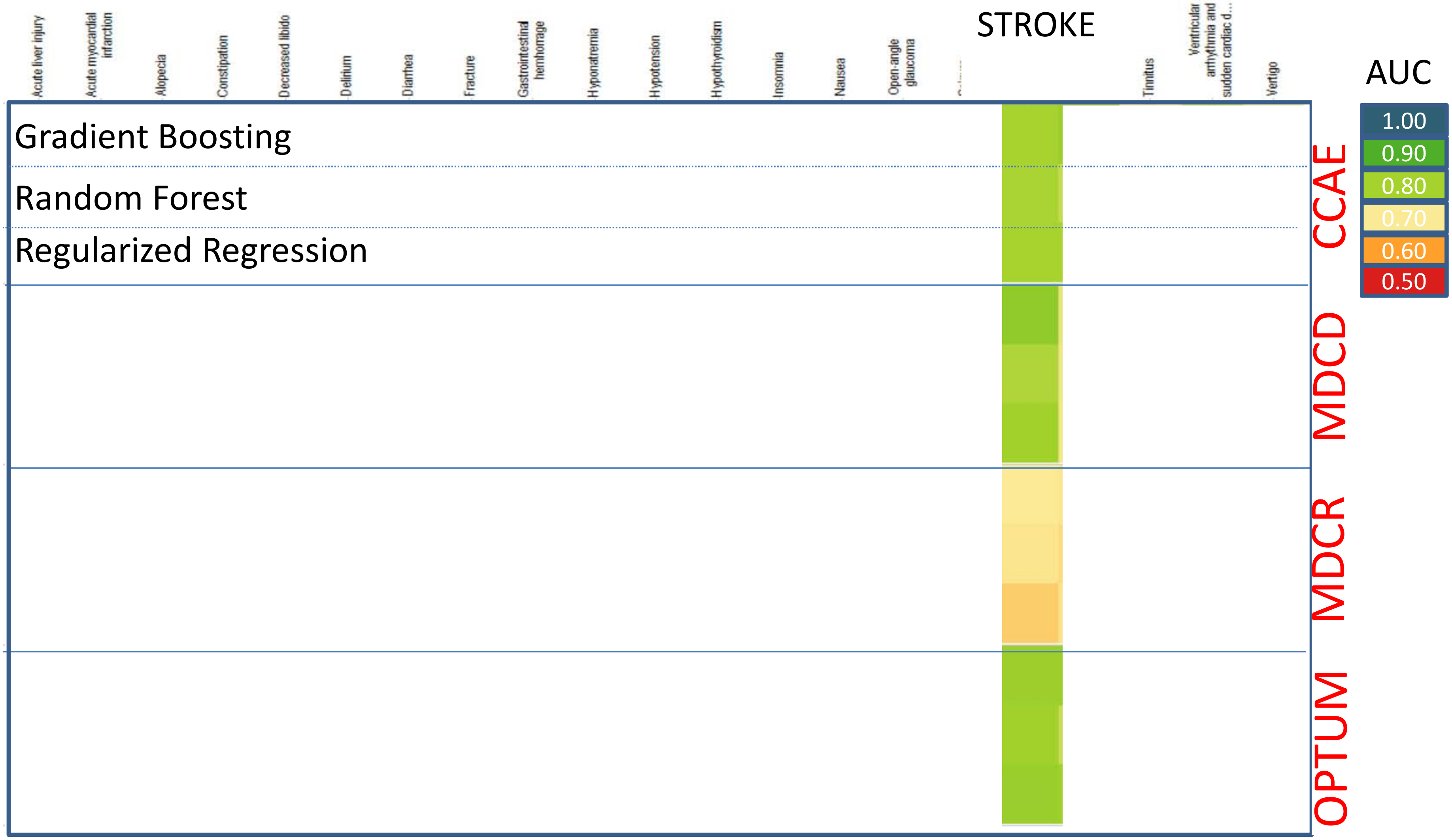
Size: value
Red: positive
Green: negative

The OHDSI approach lets the model choose from all conditions and drugs

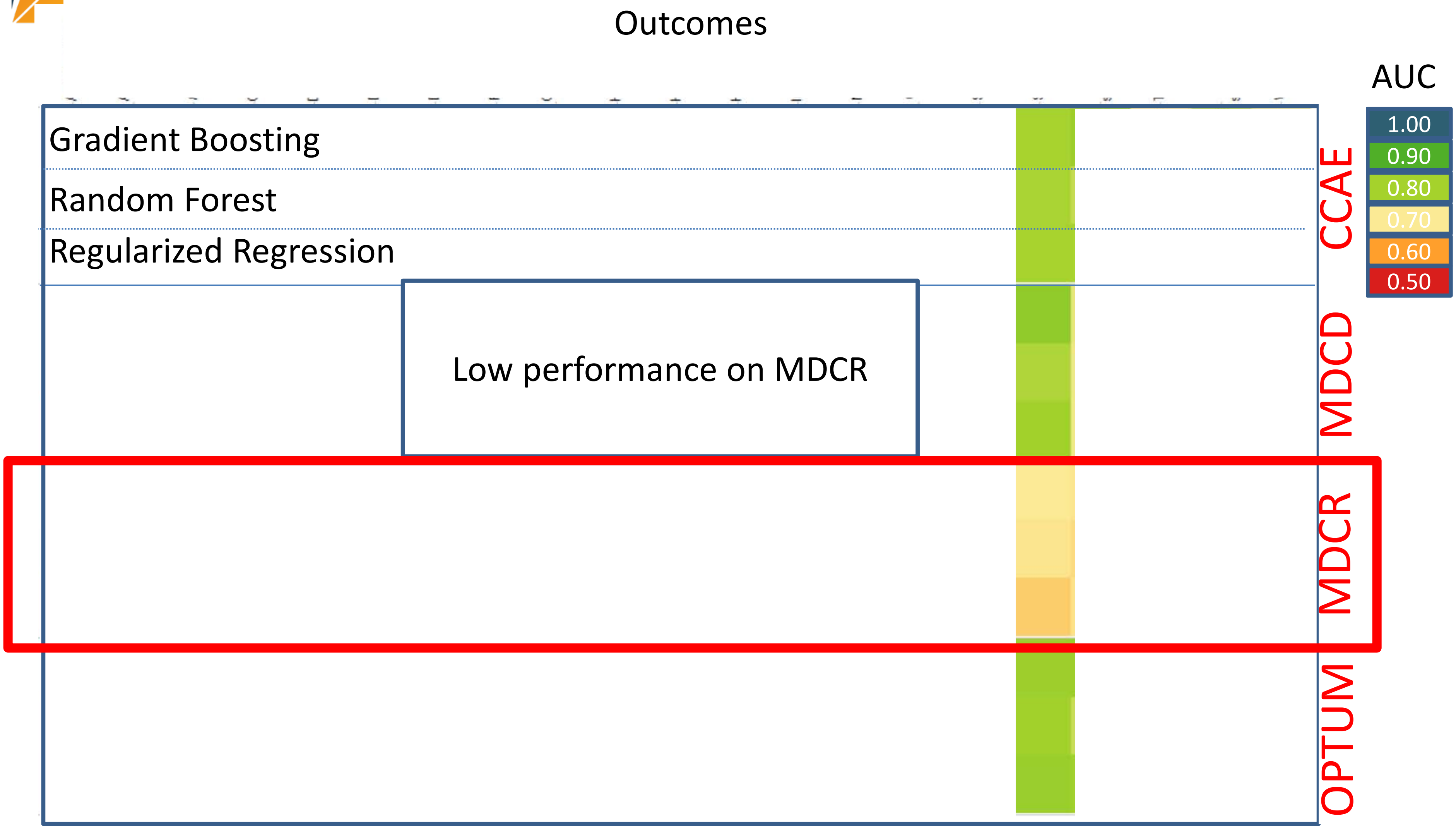
247 variables out of 16900 including:

1. all the CHADS2 markers
2. plus some other variables that make clinical sense (ex: brain cancer, smoking)
3. plus some other variables that warrant further exploration (ex: antiepileptic, COPD

Model Discrimination Stroke

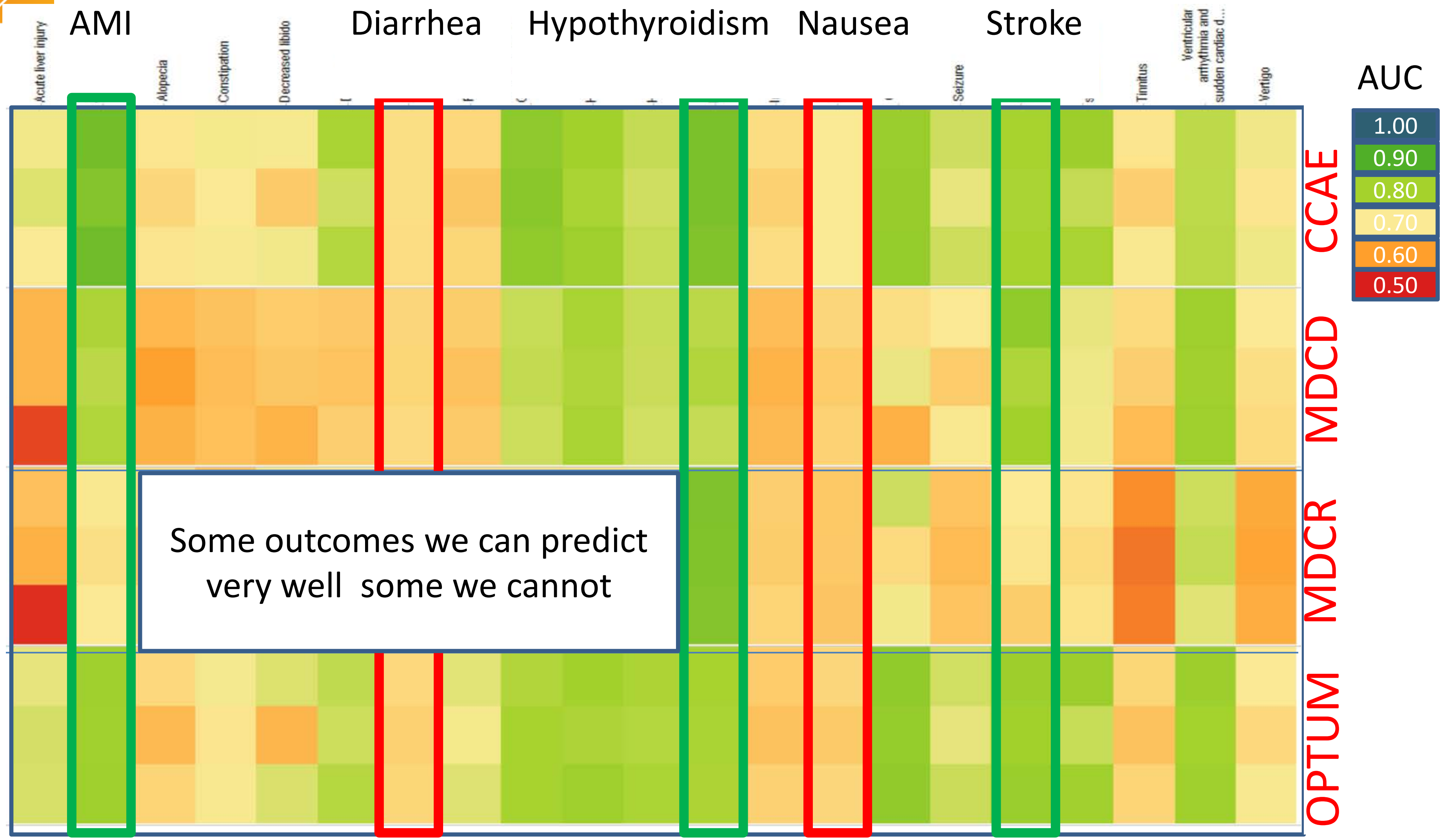


Model Discrimination



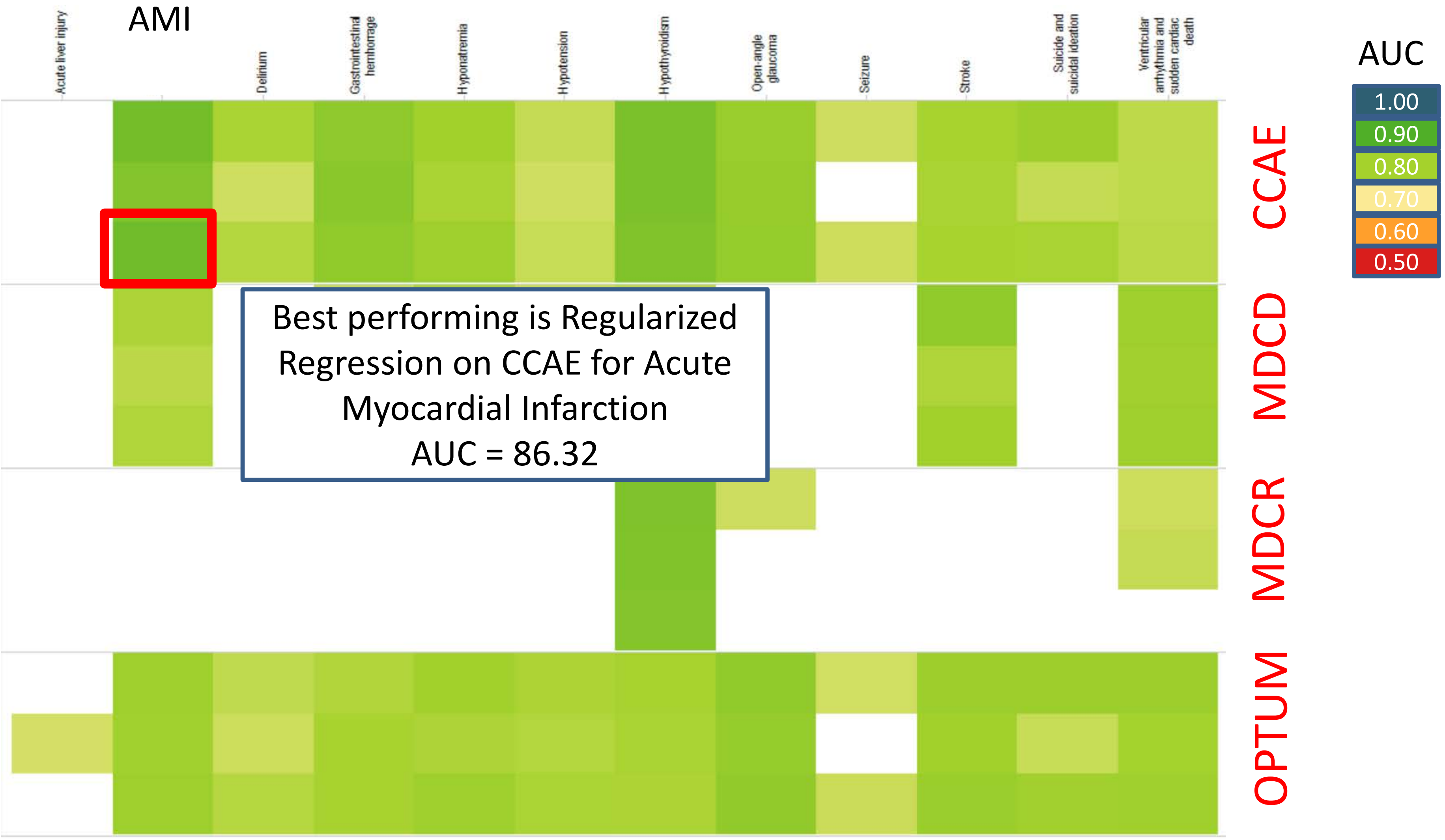
Low performance on MDCR

Model Discrimination



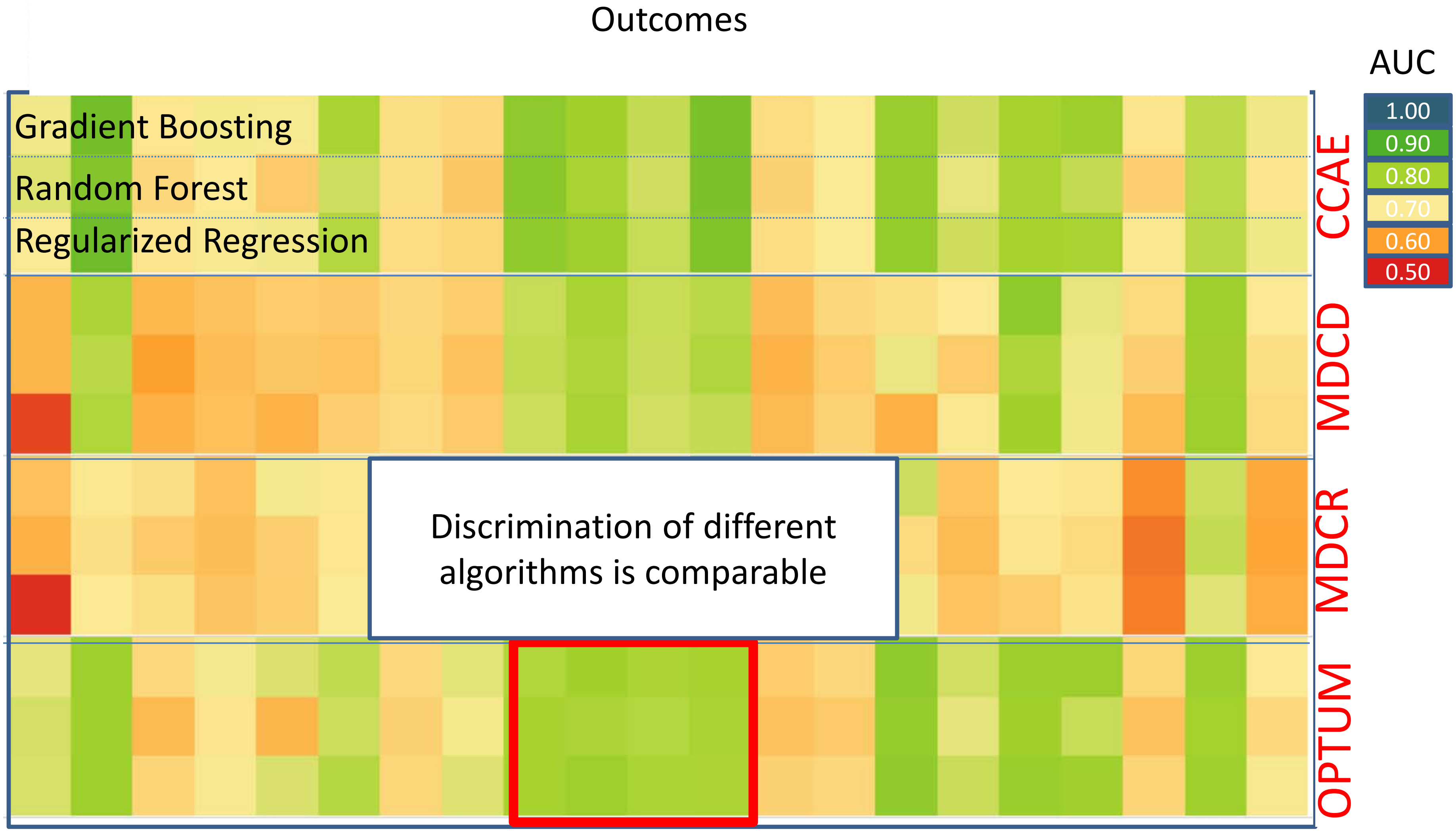
Some outcomes we can predict very well some we cannot

Outcomes with AUC > 0.75



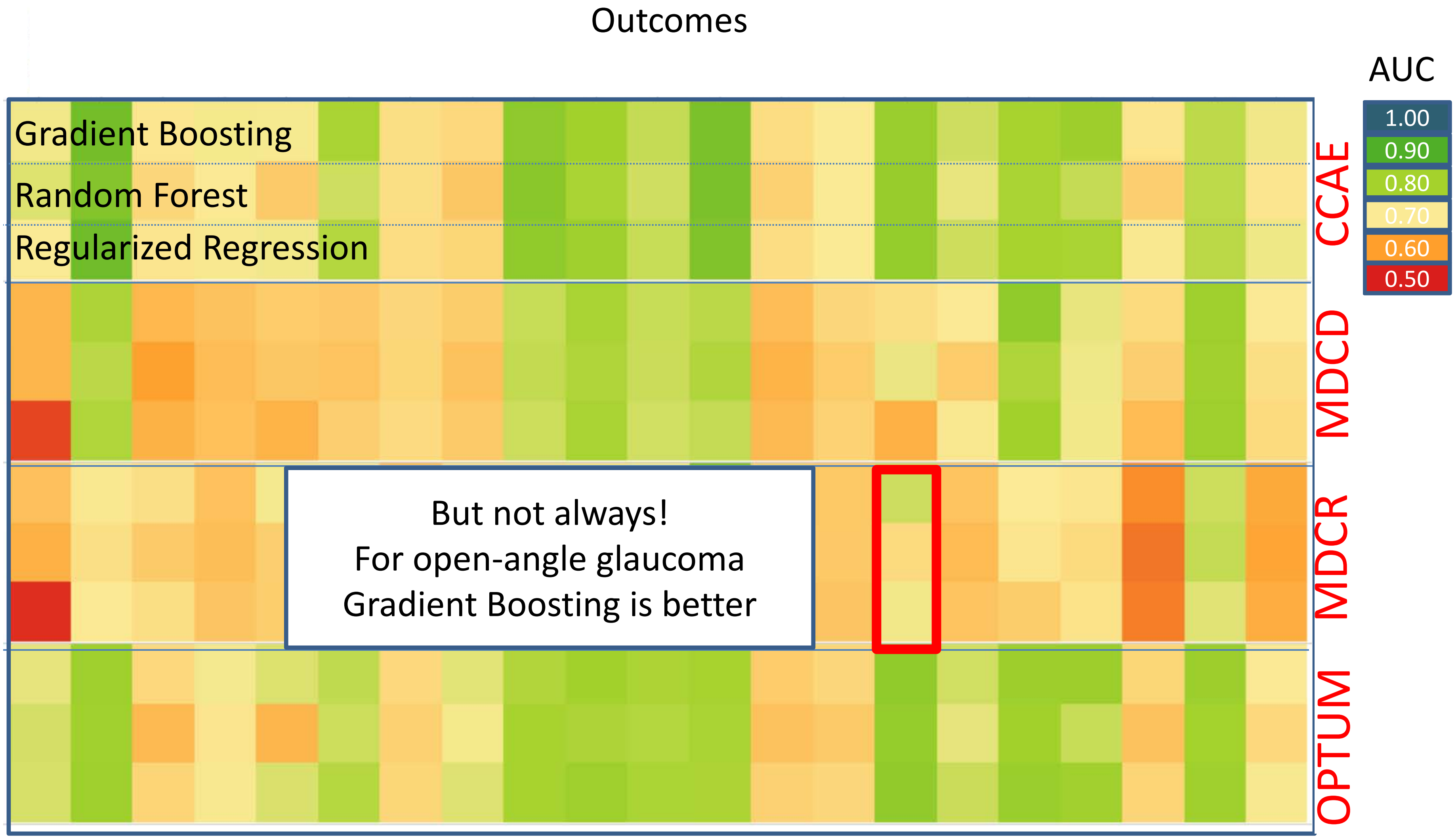
Best performing is Regularized Regression on CCAE for Acute Myocardial Infarction
AUC = 86.32

Model Discrimination



Discrimination of different algorithms is comparable

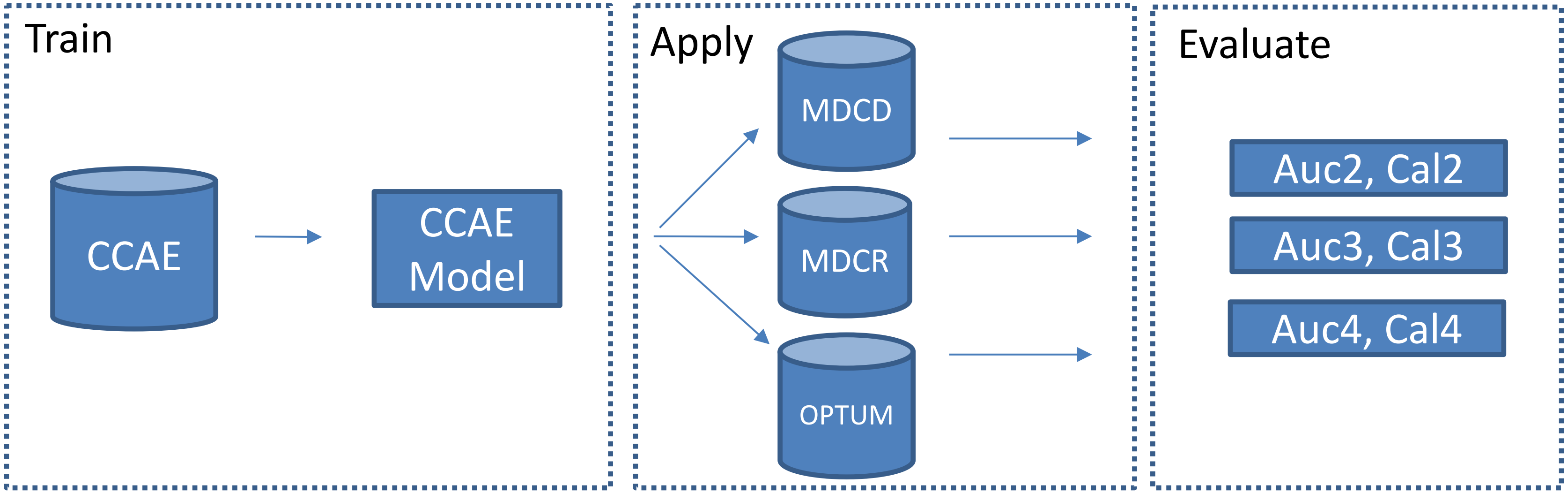
Model Discrimination



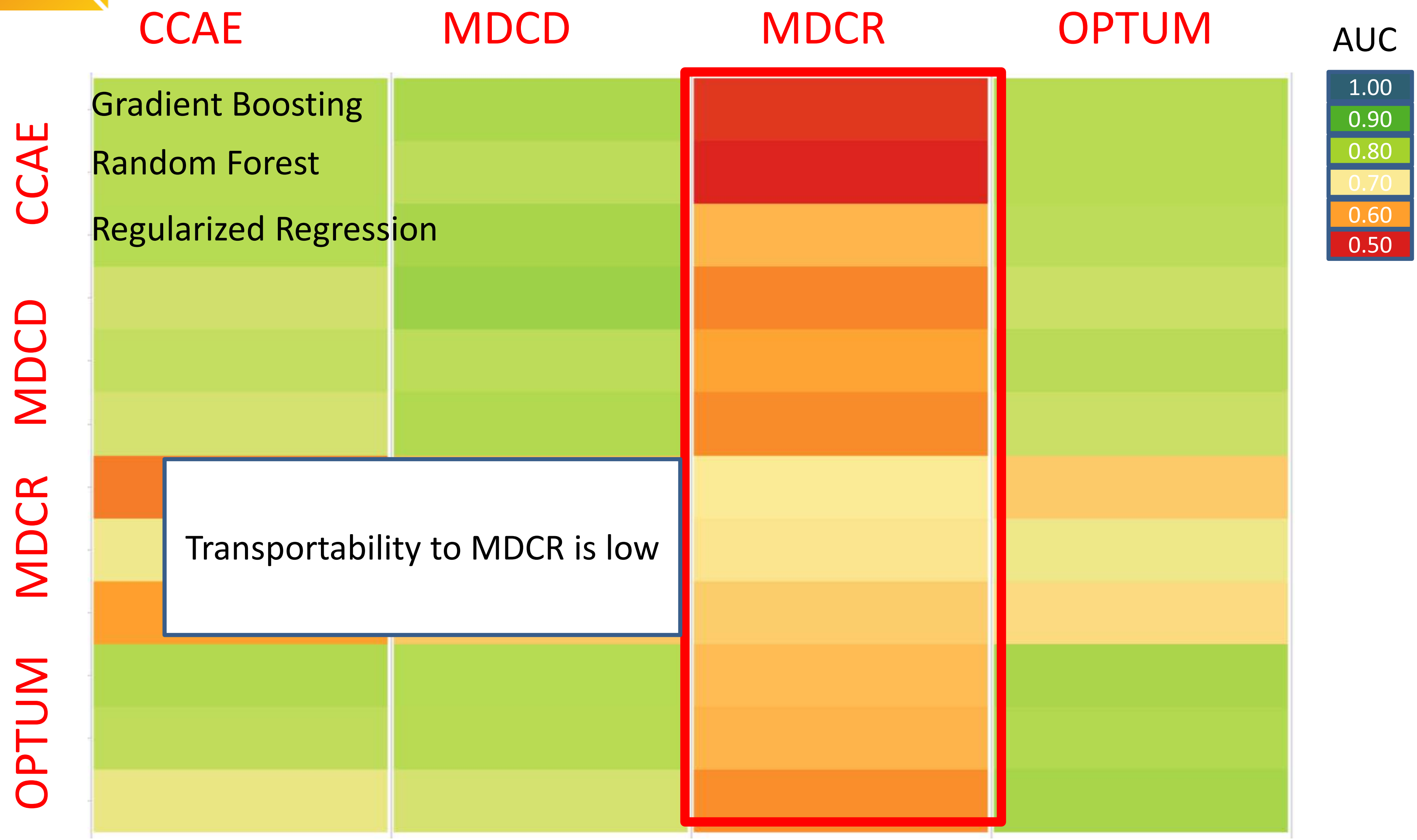
But not always!
For open-angle glaucoma
Gradient Boosting is better

Transportability Assessment

How well do the models perform on other databases?

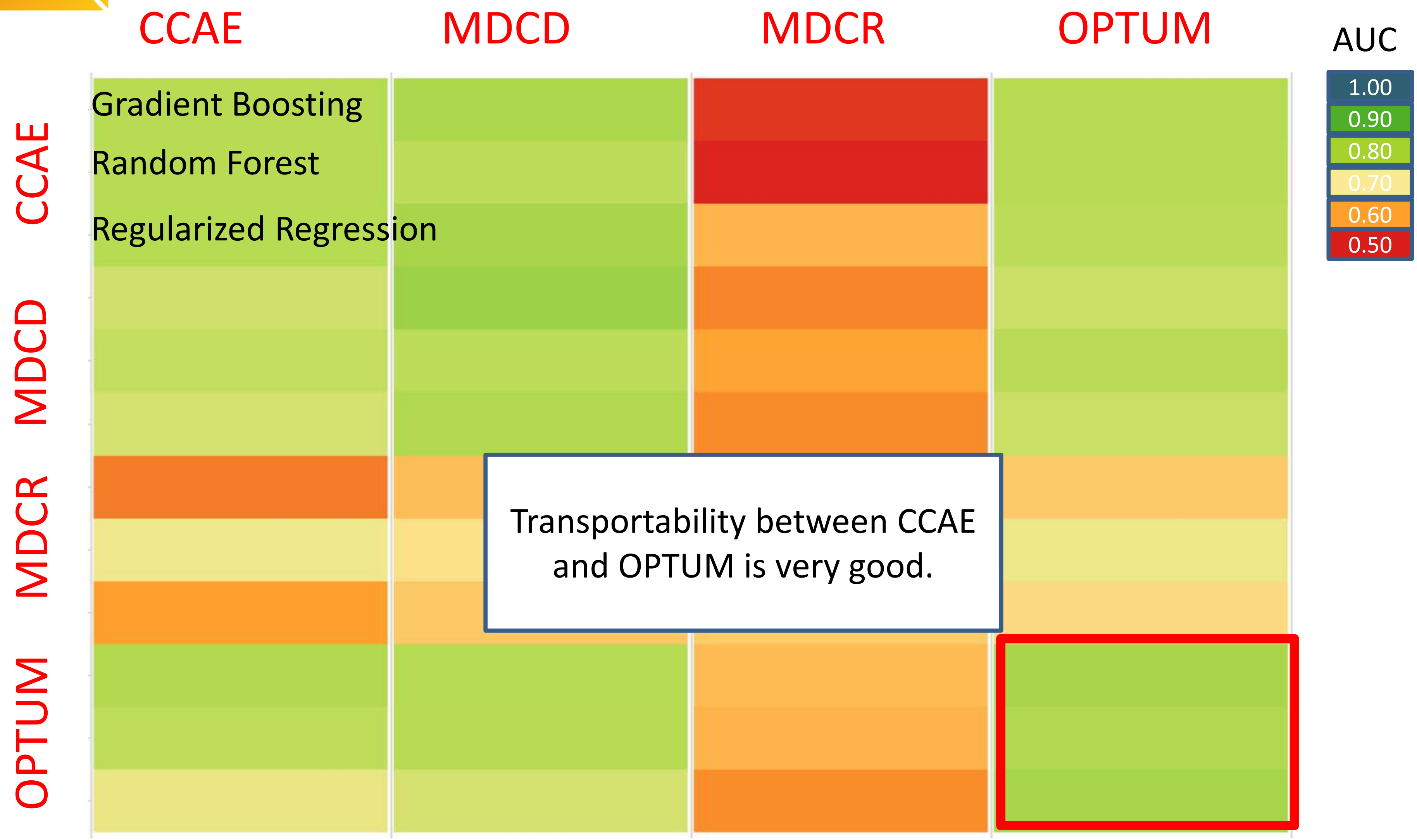


Transportability Assessment Stroke



Transportability to MDCR is low

Transportability Assessment Stroke



Transportability between CCAE and OPTUM is very good.

What did we achieve so far?

We showed it is feasible to develop large-scale predictive models for all databases converted to the OMOP CDM. This can now be done for any cohort at risk, outcome, and time at risk.

Continuation of the PLP Journey

Scale up

- Increase the number of database
- Increase the number of cohorts at risk
- Increase the number of outcomes

Method Research

- Performance
- Speed
- Transportability
- Temporal information
- Textual information
- …

Clinical impact for the patient

- How to assess?

We need you!

- We need contributions from many disciplines: clinicians, statisticians, machine learning experts, data custodians etc.
- Join the large-scale patient prediction study.
- Join the Patient-Level Prediction workgroup: http://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:patient-level_prediction

p.rijnbeek@erasmusmc.nl
jreps@its.jnj.com

Posters and Demo

- In the afternoon visit the demo of the Patient-Level Prediction R-package

- Visit our posters:
  1. **Best Practices for Patient-Level Prediction in OHDSI**
  2. **Utilizing the OHDSI collaborative network for large-scale prognostic model validation**

Join the journey!