OHDSI Tutorial: Patient-level predictive modelling in observational healthcare data

Faculty:

Peter Rijnbeek (Erasmus MC)

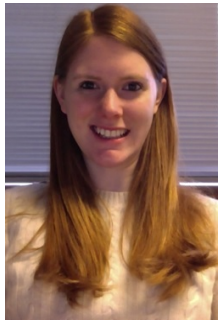
Ross Williams (Erasmus MC)

Patrick Ryan (Janssen Research and Development)

Jenna Reps (Janssen Research and Development)

The journey toward Patient-Level Prediction

Faculty

Peter Rijnbeek Erasmus MC	Ross Williams Erasmus MC	Jenna Reps Janssen R&D	Patrick Ryan Janssen R&D
			

Slides can be found on our Github:

https://github.com/OHDSI/Tutorial-PLP

Welcome to the Patient-Level Prediction Tutorial

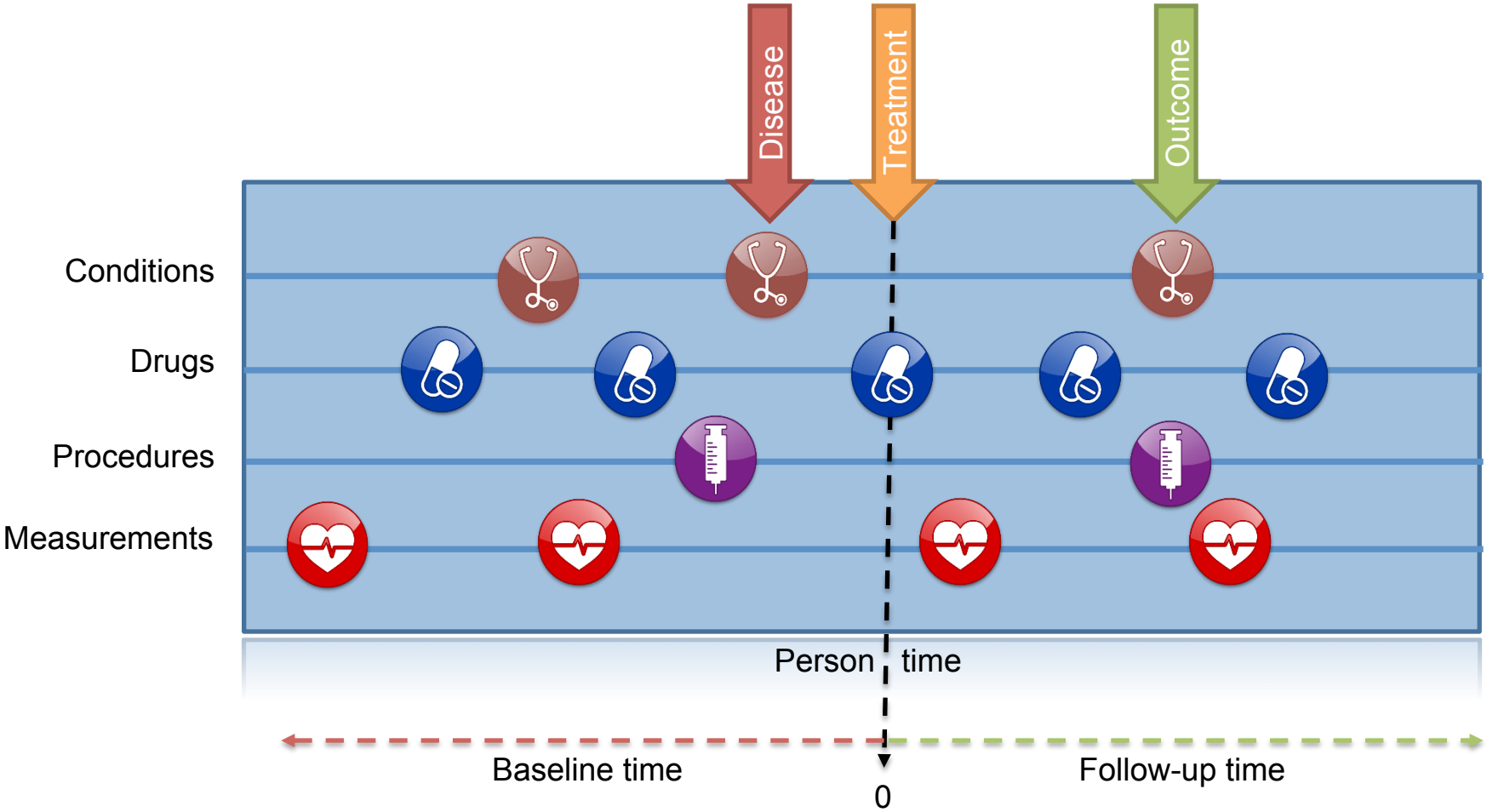
Peter Rijnbeek

OHDSI's Mission

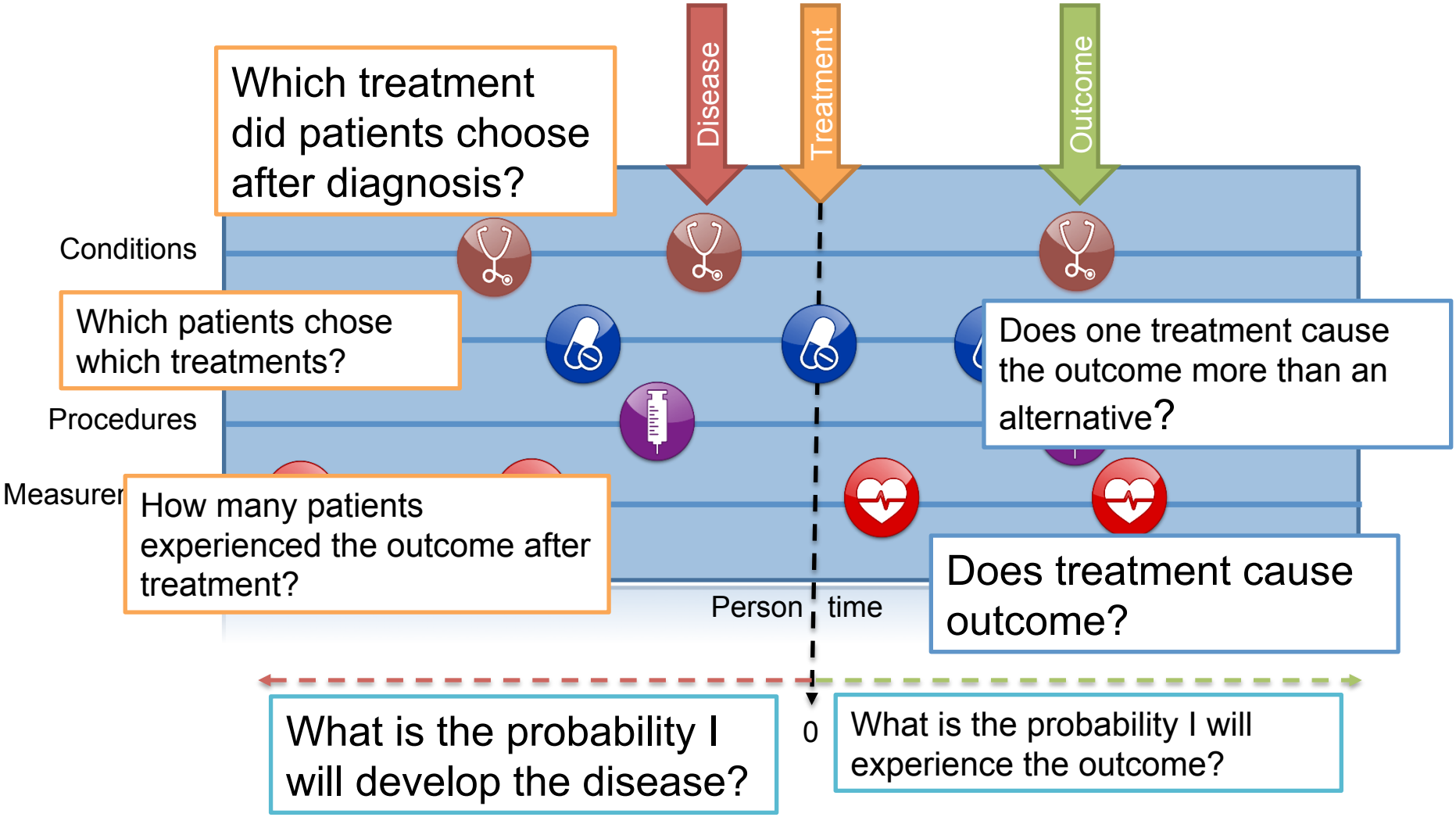
To improve health, by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care.

Hripcsak G, et al. (2015) Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers. Stud Health Technol Inform 216:574–578.

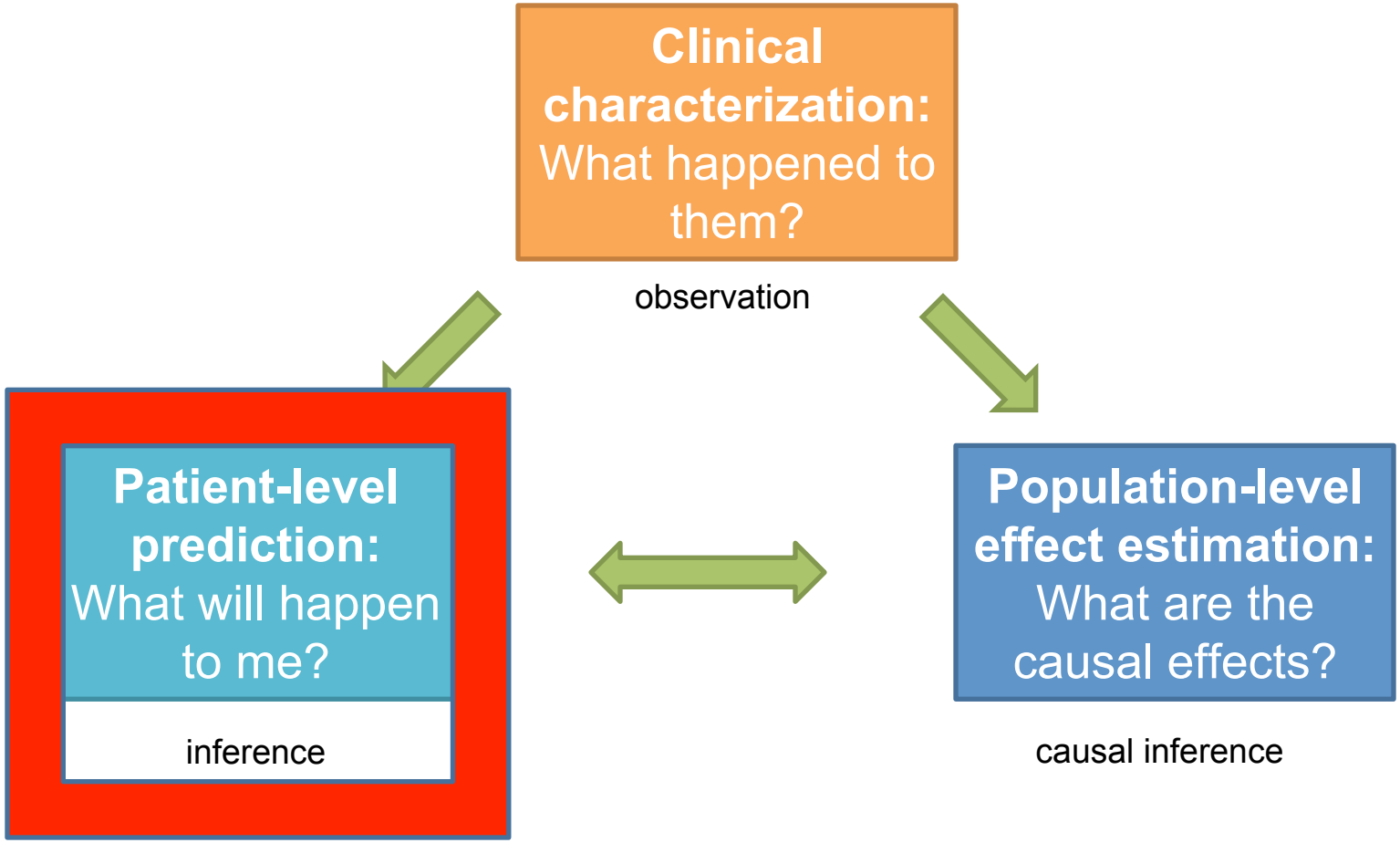
A caricature of the patient journey



Questions asked across the patient journey



Complementary evidence to inform the patient journey



Today's Agenda

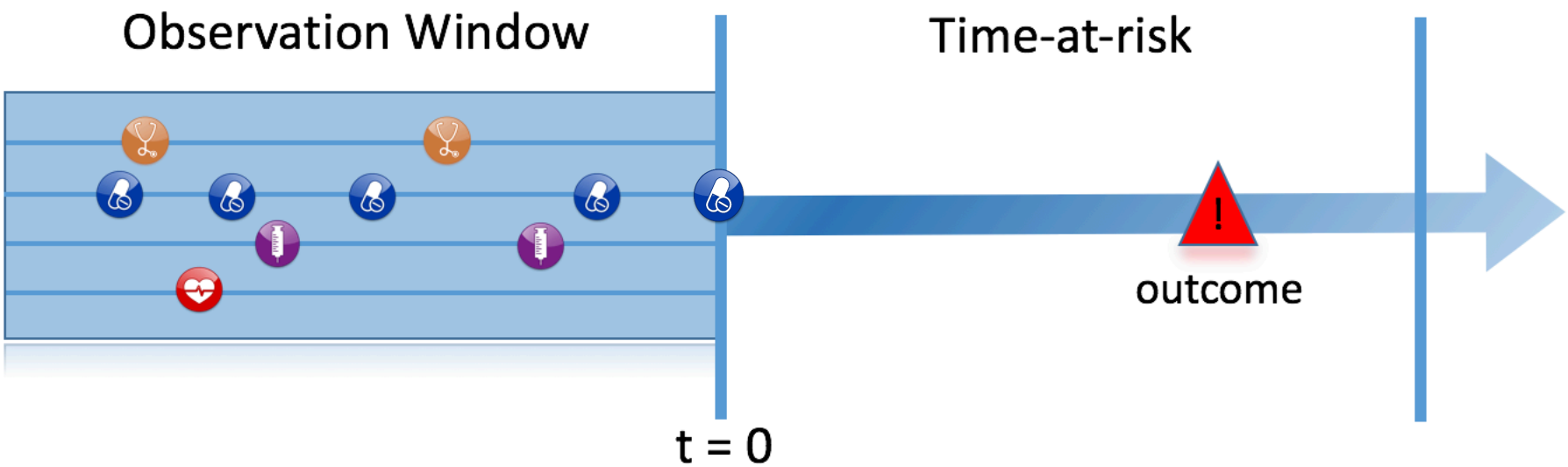
Time	Topic
8:45 – 9:00	Get settled, get laptops ready
9:00 – 10:00	Exercise: Selection of prediction problem
10:00 – 10:45	Presentation: What is Patient-Level Prediction
10:45 – 11:00	Break
11:00 – 11:45	Presentation: Learning the OHDSI Patient-Level Prediction Framework
11:45 – 12:30	Presentation: Overview of the TRIPOD Statement Exercise: Applying TRIPOD to CHADS2
12:30 – 13:15	Lunch
13:15 – 15:15	Guided tour through implementing patient-level prediction
15:15 – 15:30	Break
15:30 – 16:45	Exercise: Design and implement your own patient-level prediction
16:45 – 17:00	Lessons Learned and Feedback

Selection of prediction problem

Patrick Ryan

OHDSI
Observational Health Data Sciences and Informatics

Prediction Problem Definition

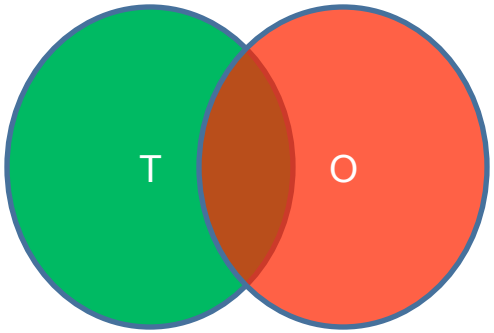


Among a target population (T), we aim to predict which patients at a defined moment in time (t=0) will experience some outcome (O) during a time-at-risk Prediction is done using only information about the patients in an observation window prior to that moment in time.

What are the key inputs to a patient-level prediction study?

Input parameter	Design choice
Target cohort (T)	
Outcome cohort (O)	
Time-at-risk	
Model development -which algorithm(s)? -which parameters? -which covariates?	

We extract data for the patients in the Target Cohort (T) of which some will experience the outcome (O) in T



Types of prediction problems in healthcare

Type	Structure	Example
Disease onset and progression	Amongst patients who are newly diagnosed with **<insert your favorite disease>**, which patients will go on to have **<another disease or related complication>** within **<time horizon from diagnosis>**?	Among newly diagnosed AFib patients, which will go onto to have ischemic stroke in next 3 years?
Treatment choice	Amongst patients with **<indicated disease> who are treated with either <treatment 1> or <treatment 2>**, which patients were treated with **<treatment 1>** (**on day 0**)?	Among AFib patients who took either warfarin or rivaroxaban, which patients got warfarin? (as defined for propensity score model)
Treatment response	Amongst patients who are new users of **<insert your favorite chronically-used drug>**, which patients will **<insert desired effect>** in **<time window>**?	Which patients with T2DM who start on metformin stay on metformin after 3 years?
Treatment safety	Amongst patients who are new users of **<insert your favorite drug>,** which patients will experience **<insert your favorite known adverse event from the drug profile>** within **<time horizon following exposure start>**?	Among new users of warfarin, which patients will have GI bleed in 1 year?
Treatment adherence	Amongst patients who are new users of **<insert your favorite chronically-used drug>**, which patients will achieve **<adherence metric threshold>** at **<time horizon>**?	Which patients with T2DM who start on metformin achieve >=80% proportion of days covered at 1 year?

Types of prediction problems in healthcare

Type	Structure
Disease onset and progression	Amongst patients who are newly diagnosed with **<insert your favorite disease>**, which patients will go on to have **<another disease or related complication>** within **<time horizon from diagnosis>**? Among newly diagnosed AFib patients, which will go onto to have ischemic stroke in next 3 years?

Types of prediction problems in healthcare

Type	Structure
Treatment choice	Amongst patients with **<indicated disease> who are treated with either <treatment 1> or <treatment 2>**, which patients were treated with **<treatment 1>** (**on day 0**)? Among AFib patients who took either warfarin or rivaroxaban, which patients got warfarin? (as defined for propensity score model)

Types of prediction problems in healthcare

Type	Structure
Treatment response	Amongst patients who are new users of **<insert your favorite chronically-used drug>**, which patients will **<insert desired effect>** in **<time window>**? Which patients with T2DM who start on metformin stay on metformin after 3 years?

Types of prediction problems in healthcare

Type	Structure
Treatment safety	Amongst patients who are new users of **<insert your favorite drug>,** which patients will experience **<insert your favorite known adverse event from the drug profile>** within **<time horizon following exposure start>**? Among new users of warfarin, which patients will have GI bleed in 1 year?

Types of prediction problems in healthcare

Type	Structure
Treatment adherence	Amongst patients who are new users of **<insert your favorite chronically-used drug>**, which patients will achieve **<adherence metric threshold>** at **<time horizon>**? Which patients with T2DM who start on metformin achieve >=80% proportion of days covered at 1 year?

Types of prediction problems in healthcare

Type	Structure	Example
Disease onset and progression	Amongst patients who are newly diagnosed with **<insert your favorite disease>**, which patients will go on to have **<another disease or related complication>** within **<time horizon from diagnosis>**?	Among newly diagnosed AFib patients, which will go onto to have ischemic stroke in next 3 years?
Treatment choice	Amongst patients with **<indicated disease> who are treated with either <treatment 1> or <treatment 2>**, which patients were treated with **<treatment 1>** (**on day 0**)?	Among AFib patients who took either warfarin or rivaroxaban, which patients got warfarin? (as defined for propensity score model)
Treatment response	Amongst patients who are new users of **<insert your favorite chronically-used drug>**, which patients will **<insert desired effect>** in **<time window>**?	Which patients with T2DM who start on metformin stay on metformin after 3 years?
Treatment safety	Amongst patients who are new users of **<insert your favorite drug>,** which patients will experience **<insert your favorite known adverse event from the drug profile>** within **<time horizon following exposure start>**?	Among new users of warfarin, which patients will have GI bleed in 1 year?
Treatment adherence	Amongst patients who are new users of **<insert your favorite chronically-used drug>**, which patients will achieve **<adherence metric threshold>** at **<time horizon>**?	Which patients with T2DM who start on metformin achieve >=80% proportion of days covered at 1 year?

What is your prediction problem?

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

OHDSI Patient-Level Prediction Design Exercise

Define your prediction problem:

What's your T? (Target cohort)	
What's your O? (Outcome cohort)	
What's your Time-At-Risk?	
What's your Model specification? -which model(s)? -which covariate(s)? -which parameters?	

What do you expect the model outputs will look like?

[Model outputs: covariate scatterplot, ROC, calibration]

How will you use the model outputs to meet the aim of your study?

1. Fill in your form (10 min)
2. Discuss your prediction problem in your group (20 min)
3. Select one prediction problem
4. Report back and promote your choice
5. Voting on prediction problem to implement after lunch

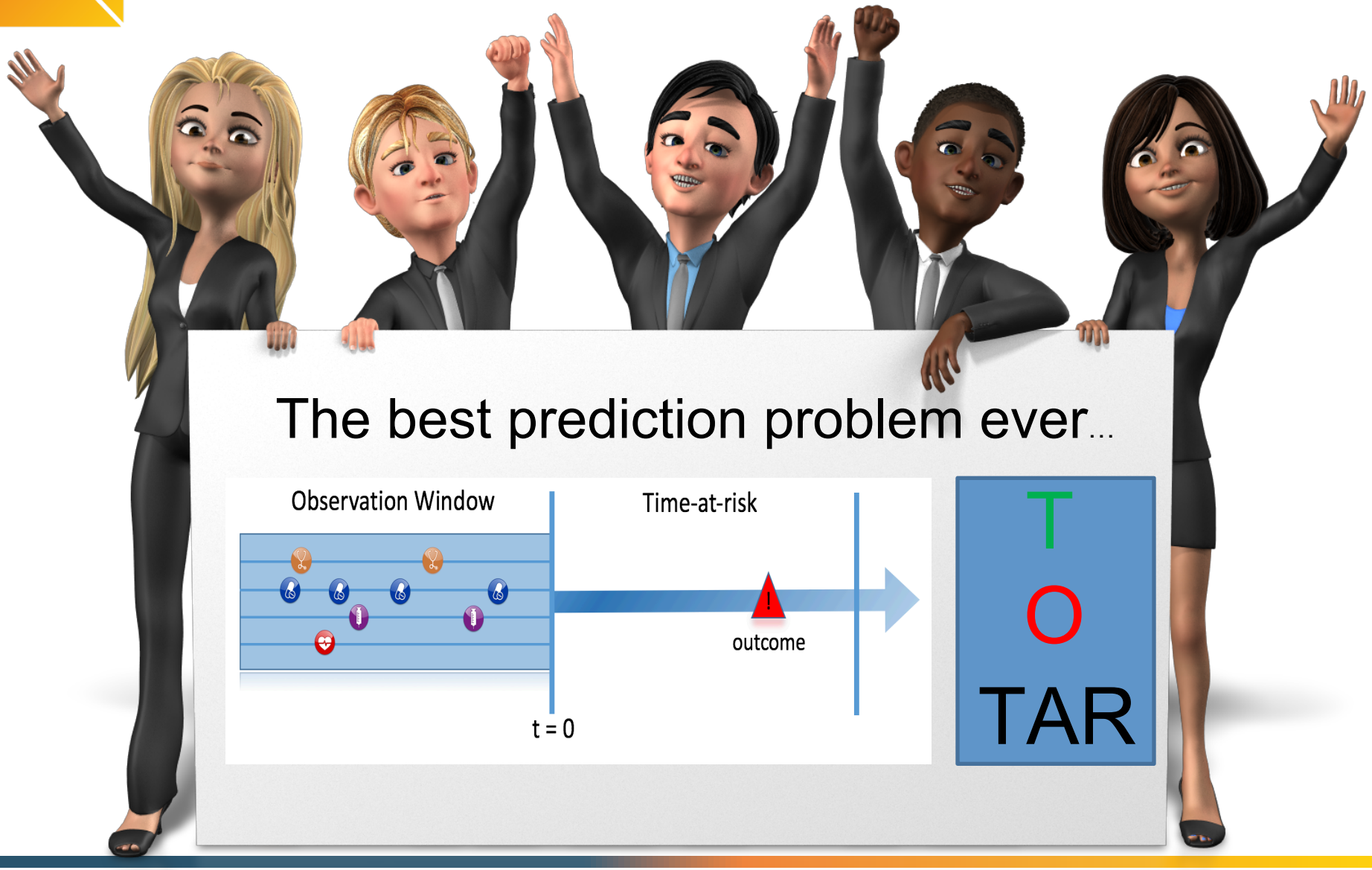
Questions?

What is your prediction problem?

You have 30 minutes for step 1 - 3

1. Fill in your form (10 min)
2. Discuss your prediction problem in your group (20 min)
3. Select one prediction problem
4. Report back and promote your choice
5. Voting on prediction problem to implement after lunch

Group Discussion



Today's Agenda

Time	Topic
8:45 – 9:00	Get settled, get laptops ready
9:00 – 10:00	Exercise: Selection of prediction problem
10:00 – 10:45	Presentation: What is Patient-Level Prediction
10:45 – 11:00	Break
11:00 – 11:45	Presentation: Learning the OHDSI Patient-Level Prediction Framework
11:45 – 12:30	Presentation: Overview of the TRIPOD Statement Exercise: Applying TRIPOD to CHADS2
12:30 – 13:15	Lunch
13:15 – 15:15	Guided tour through implementing patient-level prediction
15:15 – 15:30	Break
15:30 – 16:45	Exercise: Design and implement your own patient-level prediction
16:45 – 17:00	Lessons Learned and Feedback

What is Patient-Level Prediction?

Peter Rijnbeek, PhD
Erasmus MC

Learning Objectives

Part 1: Learn what a patient-level prediction model is?

Part 2: Understand the patient-level prediction modelling process

Part 3: Gain insights from a proof-of-concept study in depression patients

Clinicians are confronted with prediction questions on a daily basis. What options do they have?

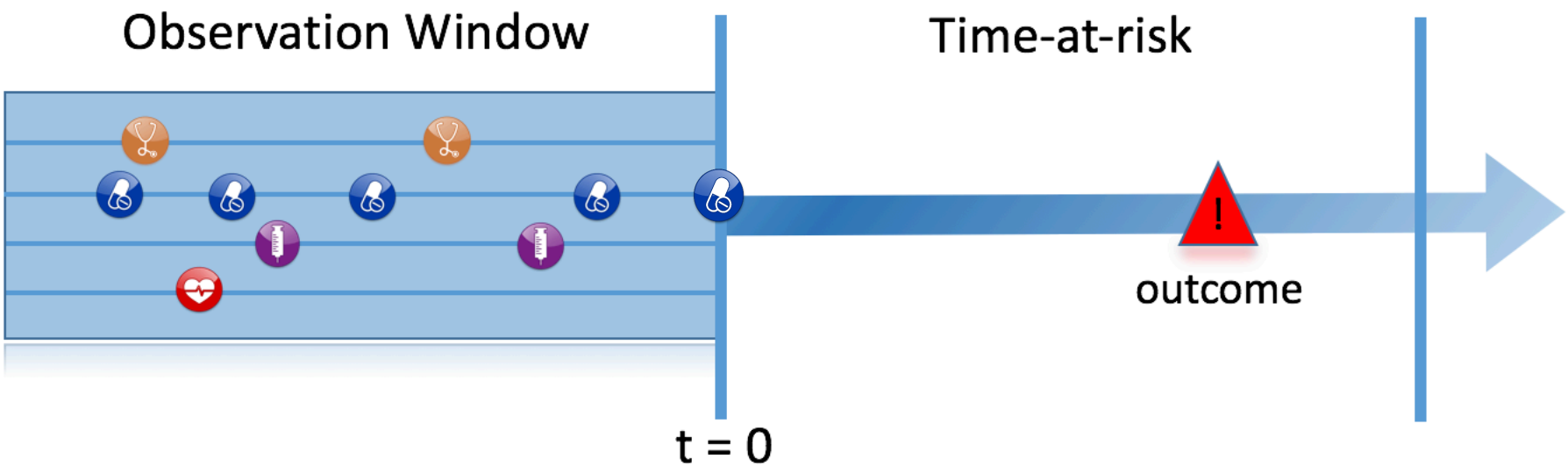
Deny ability to predict at the individual patient level

Quote an overall average to all patients

Utilize knowledge and personal experience

Provide a personalized prediction based on an advanced clinical prediction model

Problem definition

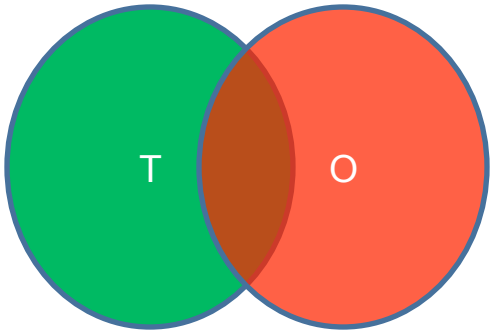


Among a target population (T), we aim to predict which patients at a defined moment in time (t=0) will experience some outcome (O) during a time-at-risk Prediction is done using only information about the patients in an observation window prior to that moment in time.

What are the key inputs to a patient-level prediction study?

Input parameter	Design choice
Target cohort (T)	
Outcome cohort (O)	
Time-at-risk	
Model specification -which model(s)? -which parameters? -which covariates?	

We extract data for the patients in the Target Cohort (T) and we select all patients that experience the outcome (O) in T



Difference between explanatory models and prediction models

People build a prediction model and make causal claims. This is not correct!

Why

Different interpretations of ”Model”

”Model” is being interpreted differently in Statistics, Epidemiology, and Data Science

- Statistics: models are used to describe data, it is more about data characterization
- Epidemiologist are trained to think about models as tests of hypotheses to perform causal inference
- Data Scientists interpret the word “model” in the context of predicting future events using the available data

It is important we understand what the difference is between explanatory modelling and predictive modelling!

Shmueli, G. 2011. Predictive Analytics in Information Systems Research. MIS Quarterly (35:3), pp. 553-57

Shmueli, G. 2010. To Explain or to Predict?, Statistical Science (25:3), pp. 289-310

Some definitions

Explanatory Model:	Theory-based statistical model for testing causal hypotheses
Explanatory Power:	Strength of the relationship in statistical model
Predictive Model:	Empirical model/algorithm for predicting new observations
Predictive Power:	Ability to accurately predict new observations

You can empirically evaluate the predictive power of explanatory model but you cannot empirically evaluate the explanatory power of a predictive model.

The best explanatory model is not necessary the best predictive model!

You do not have to understand the underlying causes in order to predict well!

Explanatory modelling versus Predictive analytics

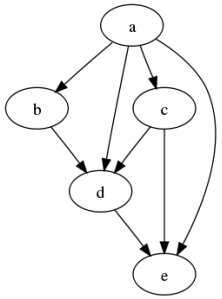


Table 1. Differences Between Explanatory Statistical Modeling and Predictive Analytics

Step	Explanatory	Predictive
Analysis Goal	Explanatory statistical models are used for testing causal hypotheses.	Predictive models are used for predicting new observations and assessing predictability levels.
Variables of Interest	Operationalized variables are used only as instruments to study the underlying conceptual constructs and the relationships between them.	The observed, measurable variables are the focus.
Model Building Optimized Function	In explanatory modeling the focus is on minimizing model bias. Main risks are type I and II errors.	In predictive modeling the focus is on minimizing the combined bias and variance. The main risk is over-fitting.
Model Building Constraints	Empirical model must be interpretable, must support statistical testing of the hypotheses of interest, must adhere to theoretical model (e.g., in terms of form, variables, specification).	Must use variables that are available at time of model deployment.
Model Evaluation	Explanatory power is measured by strength-of-fit measures and tests (e.g., R^2 and statistical significance of coefficients).	Predictive power is measured by accuracy of out-of-sample predictions.

Shmueli, G. 2011. Predictive Analytics in Information Systems Research. MIS Quarterly (35:3), pp. 553-57

Why should we avoid the term "Risk Factor"

"Risk Factor" is an ambiguous term.

A predictive model is not selecting parameters based on their explanatory power but it is **using** association to improve predictive accuracy -> **association does not equal causation**!

If your goal is to search for causal factors you should use population-level effect estimation.

If your goal is to search for association of individual parameters you should use clinical characterization.

We should avoid using the term "risk factors" and use the term predictors to make explicit that we are assessing predictive value.

How to interpret beta values in a logistic regression prediction model?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

Each beta coefficient represents the additional effect of adding that variable to the model, if the effects of all other variables in the model are already accounted for.

→ any change of the model can result in a change of all the beta coefficients

Value	Association	Causation
b = 0	Unknown	Unknown
b <> 0	Yes	Unknown
b > 0	Positively associated under the assumption that all other beta values are fixed. If the variable is correlated to any other variable the direction of the association is unknown	Unknown
b < 0	Negatively associated under the assumption that all other beta values are fixed. If the variable is correlated to any other variable the direction of the association is unknown	Unknown

Why is predictive modelling still valuable?

1. In healthcare the question “What is going to happen to me?” is often more relevant than “Why?”

2. Knowing if something is predictable or not based on the available data is valuable on its own.

Types of prediction problems in healthcare

Type	Structure	Example
Disease onset and progression	Amongst patients who are newly diagnosed with **<insert your favorite disease>**, which patients will go on to have **<another disease or related complication>** within **<time horizon from diagnosis>**?	Among newly diagnosed AFib patients, which will go onto to have ischemic stroke in next 3 years?
Treatment choice	Amongst patients with **<indicated disease> who are treated with either <treatment 1> or <treatment 2>**, which patients were treated with **<treatment 1>** (**on day 0**)?	Among AFib patients who took either warfarin or rivaroxaban, which patients got warfarin? (as defined for propensity score model)
Treatment response	Amongst patients who are new users of **<insert your favorite chronically-used drug>**, which patients will **<insert desired effect>** in **<time window>**?	Which patients with T2DM who start on metformin stay on metformin after 3 years?
Treatment safety	Amongst patients who are new users of **<insert your favorite drug>,** which patients will experience **<insert your favorite known adverse event from the drug profile>** within **<time horizon following exposure start>**?	Among new users of warfarin, which patients will have GI bleed in 1 year?
Treatment adherence	Amongst patients who are new users of **<insert your favorite chronically-used drug>**, which patients will achieve **<adherence metric threshold>** at **<time horizon>**?	Which patients with T2DM who start on metformin achieve >=80% proportion of days covered at 1 year?

Questions?

Growing interest in prediction modelling

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

Predicting Stroke in patients with atrial fibrillation

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

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.

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

CHADS2	Score
Congestive Heart Failure	1
Hypertension	1
Age $\geq$ 75	1
Diabetes	1
Stroke / TIA	2

How to define the $CHADS_2$ patient-level prediction problem?

Input parameter	Design choice
Target cohort (T)	Patients newly diagnosed with AF
Outcome cohort (O)	Stroke
Time-at-risk	1000 days
Model specification	Logistic Regression using 5 pre-selected predictors

Current status of predictive 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

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

ABSTRACT

Objective Electronic health records (EHRs) are an increasingly common data source for clinical risk prediction, presenting both unique analytic opportunities and challenges. We sought to evaluate the current state of EHR based risk prediction modeling through a systematic review of clinical prediction studies using EHR data.

Methods We searched PubMed for articles that reported on the use of an EHR to develop a risk prediction model from 2009 to 2014. Articles were extracted by two reviewers, and we abstracted information on study design, use of EHR data, model building, and performance from each publication and supplementary documentation.

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.

Goldstein BA, J Am Med Inform Assoc. 2016.

Current status of predictive 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

OHDSI 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

Part 2: How to build and validate a prediction model?

OHDSI

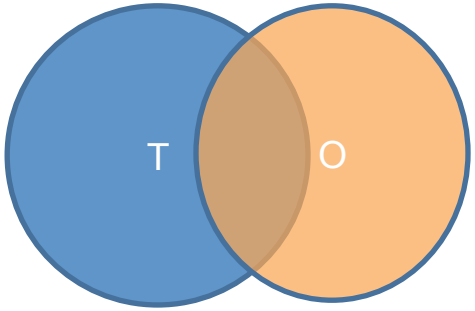
Observational Health Data Sciences and Informatics

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



We extract data for the patients in the Target Cohort (T) and we select all patients that experience the outcome (O)

The Target Cohort (T) and Outcome Cohort (O) can be defined using ATLAS or custom code (see later today).

For model development all outcomes (O) of patients in the Target Cohort (T) are used.

Prediction Model Development

Problem Definition → **Data Extraction** → Training → Internal Validation → External Validation → Dissemination

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.

This allows for specification of the candidate predictors and time windows.

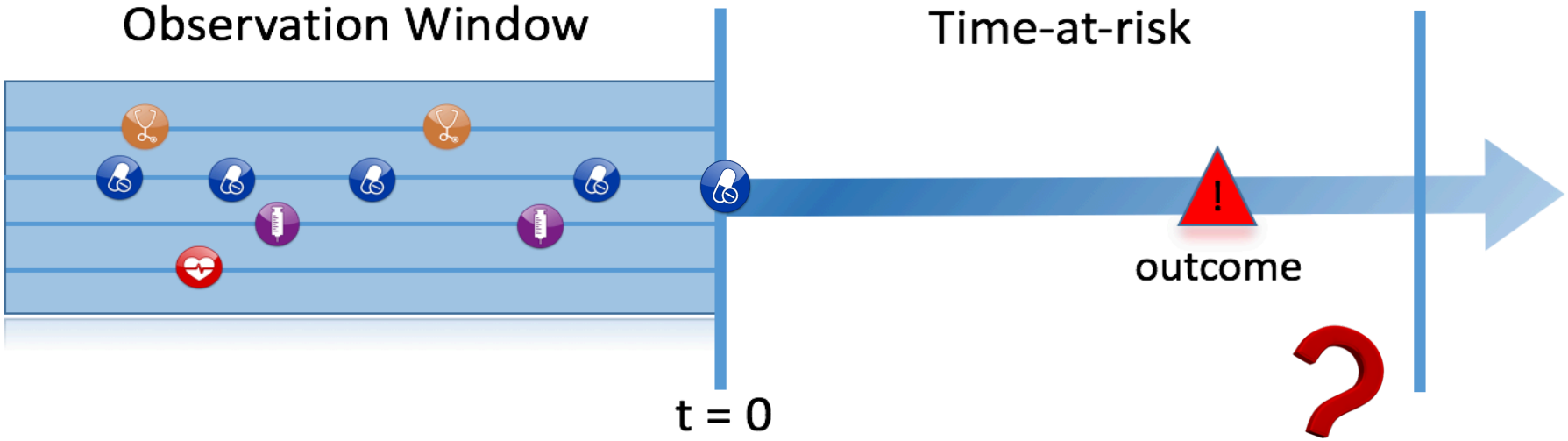
Prediction Model Development

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)

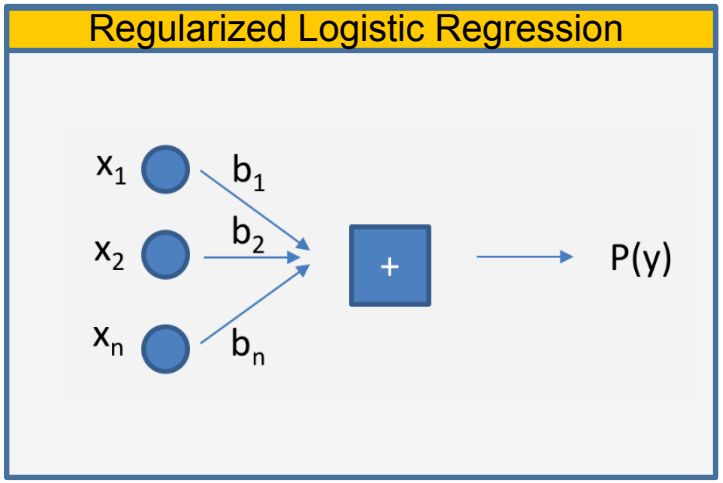
Model Training

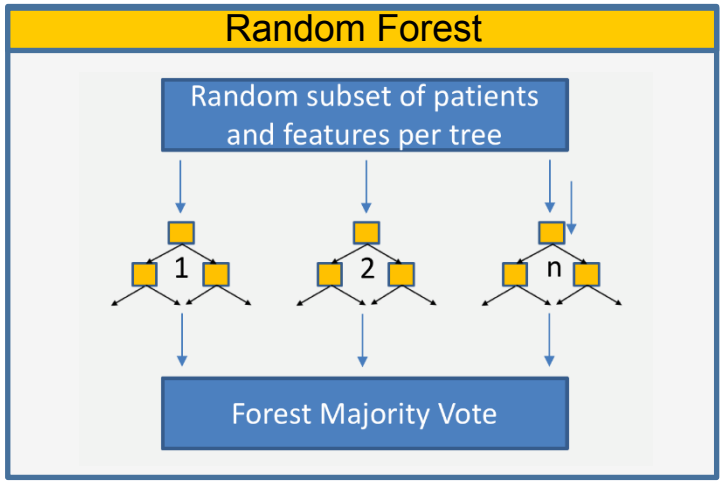


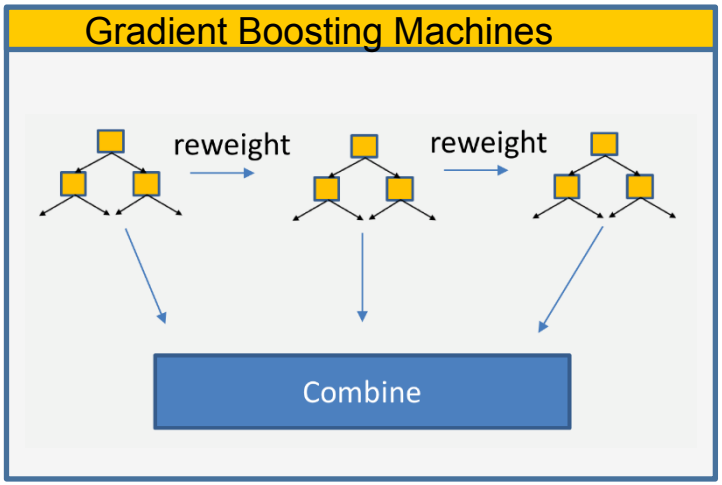
1. Which models?

2. How to evaluate the model?

Models and Algorithms







Many other models for example:

K-nearest neighbors
Naïve Bayes
Decision Tree
Adaboost
Neural Network
Deep Learning
Etc.

Model selection is an empirical process

The **"No Free Lunch" theorem** states that there is not one model that works best for every problem. The assumptions of a great model for one problem may not hold for another problem.

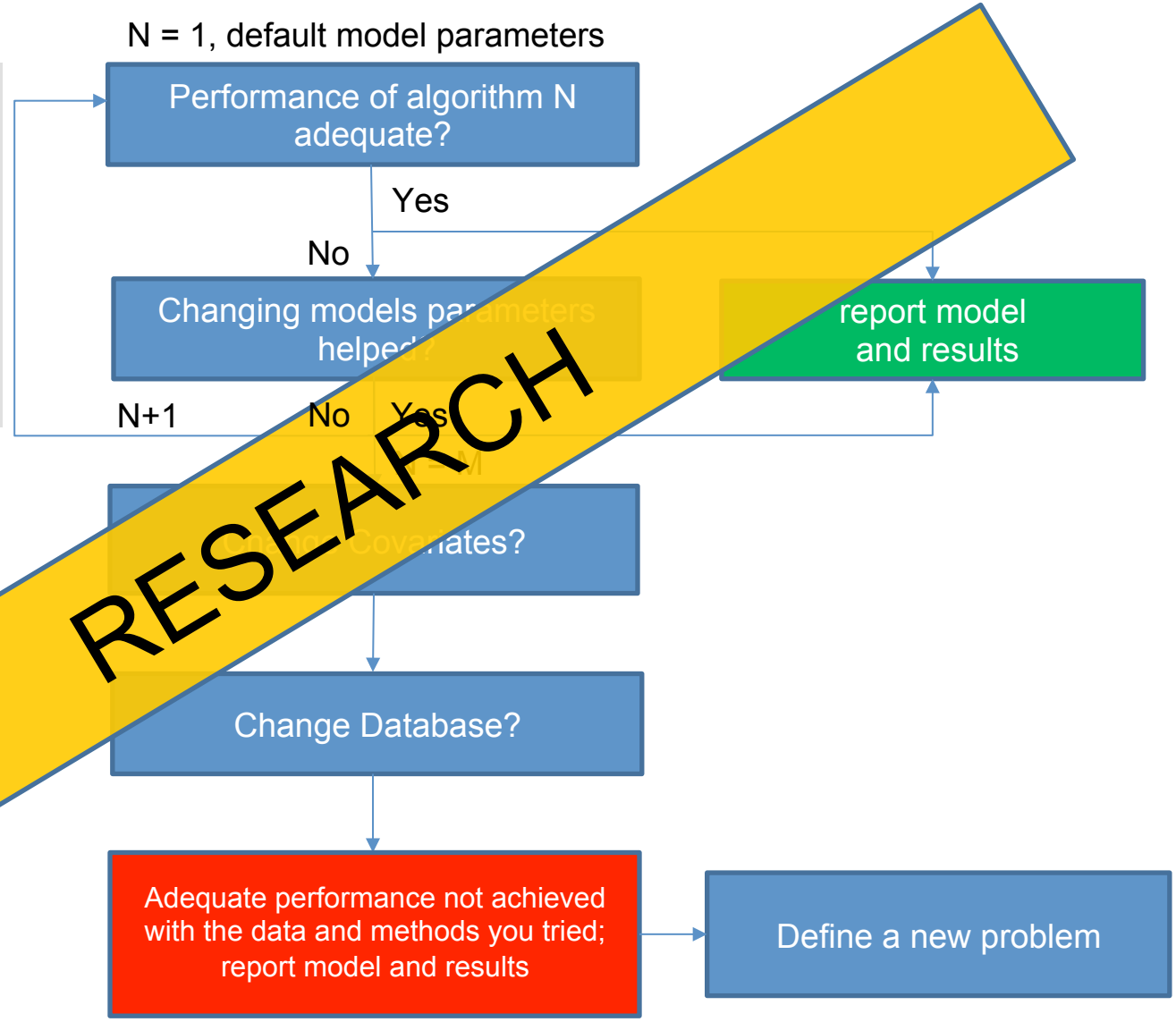
It is common in machine learning to try multiple models and find one that works best for that particular problem.

see no-free-lunch.org

OHDSI Model Selection Strategy

Suggested ordering of available algorithms in PLP package

N	Algorithms
1.	Lasso Logistic Regression
2.	Random Forest
3.	Gradient Boosting Machine
4.	Neural Network
5.	KNN
M.	…



RESEARCH

Patient-Level Prediction Roadmap

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

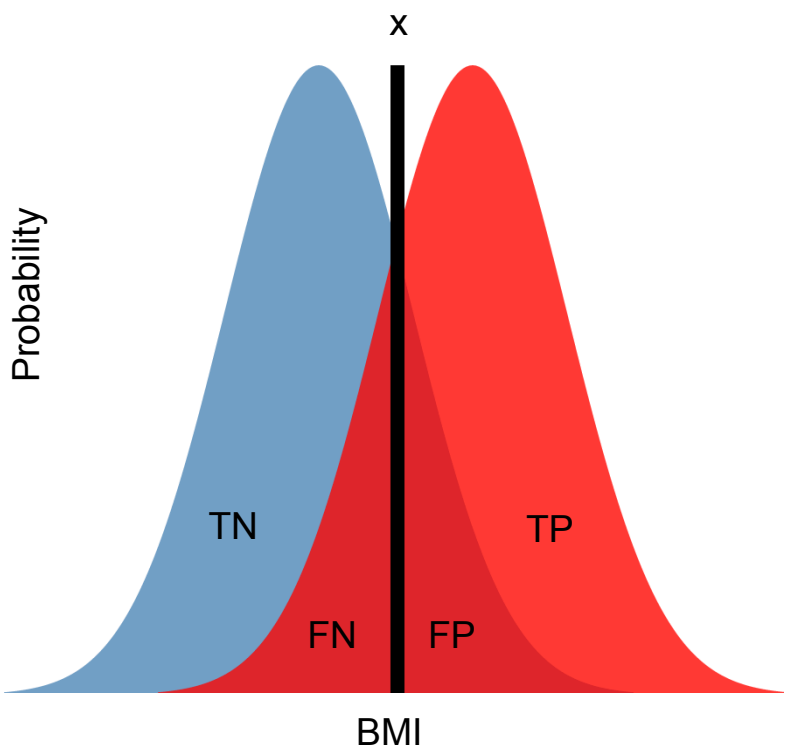
Calibration: estimated probabilities are close to the observed frequency

How to assess discrimination?

Suppose our classifier is simply BMI > x.

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

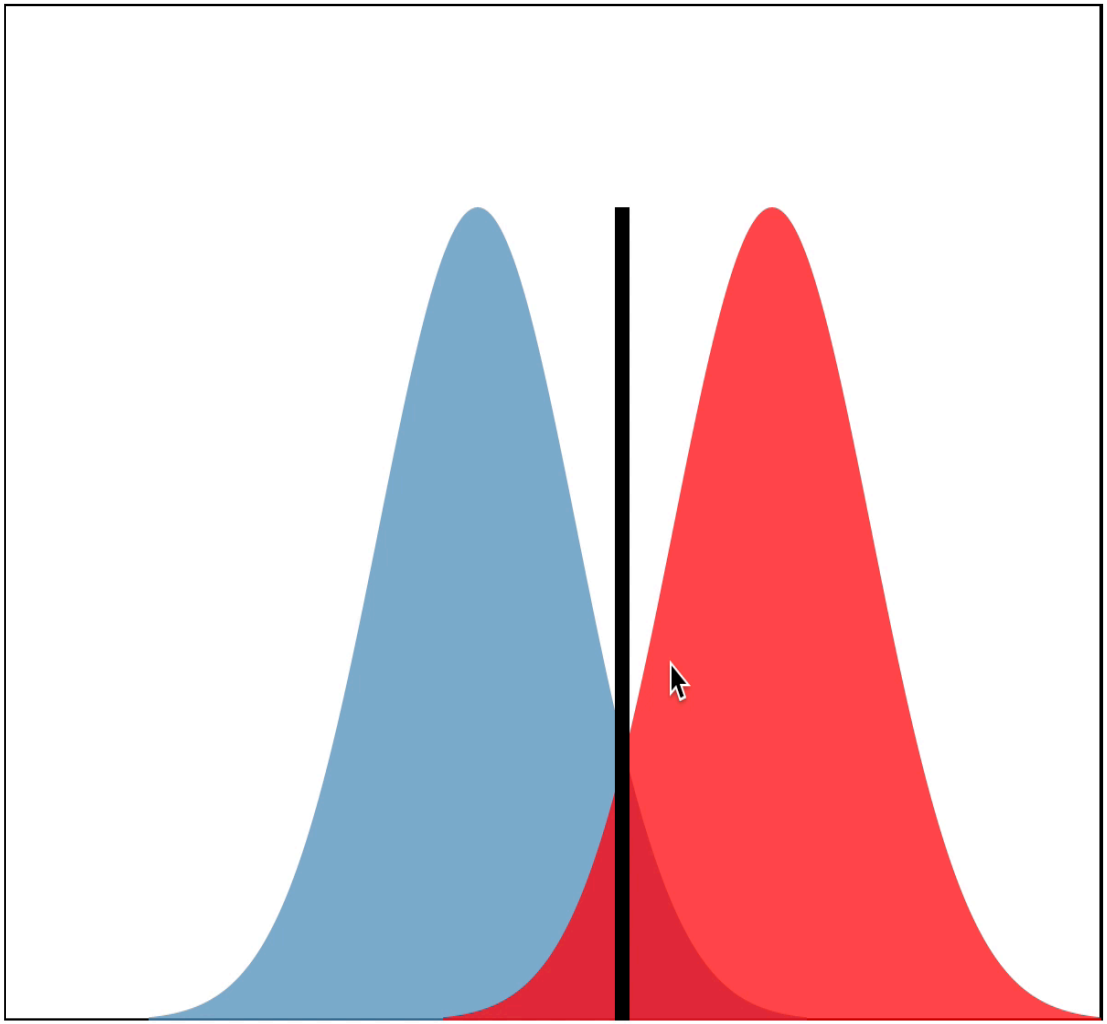
The choice of X then determines how sensitive or specific our algorithm is.



		Predicted	
		1	0
Observed	1		
	0		

True Positive Rate (TPR) = TP / (TP + FN)
False Positive Rate (FPR) = FP / (FP + TN)

Receiver Operator Characteristic (ROC) curve



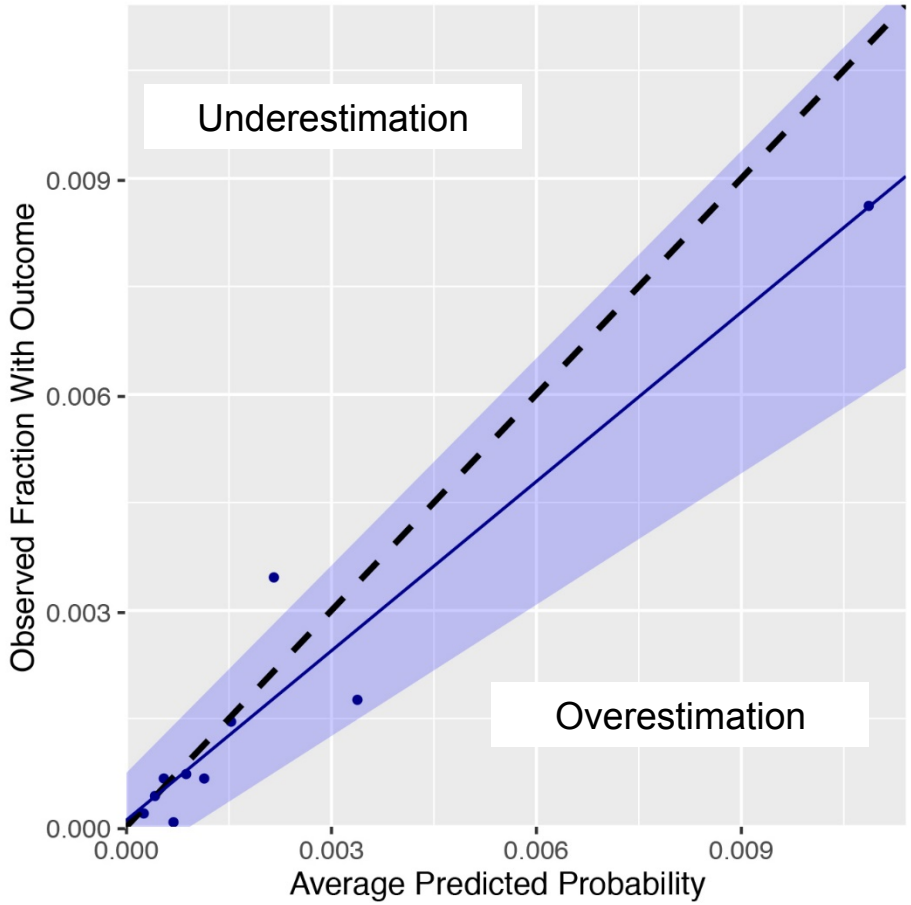
Discrimination: Area under curve (AUC)

Calibration

- Agreement between observed and predicted risk
- We want a model that has good calibration across the range of predictions (not just on average)
- A model is well calibrated if for every 100 individuals given a risk of p% close to p have the event.
- For example, if we predict a 12% risk that an atrial fibrillation patient will have a stroke within 365 days, the observed proportion should be approx. 12 strokes per 100 patients

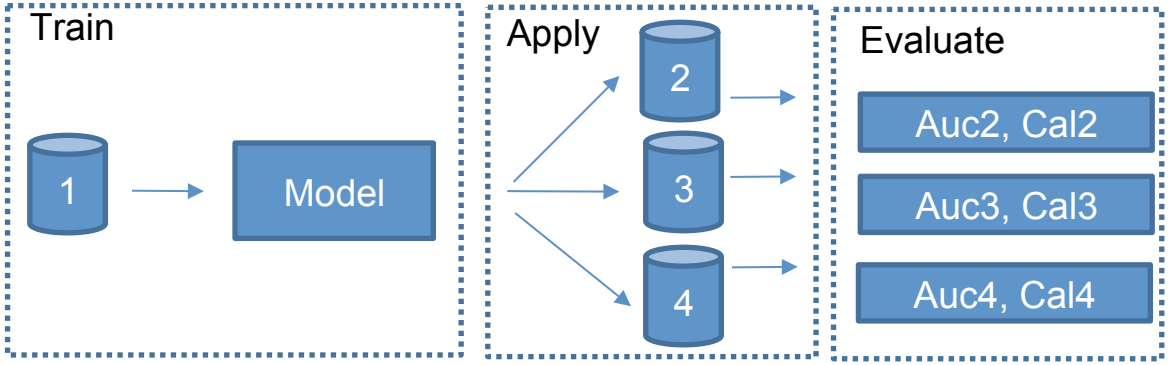
Calibration Assessment

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

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

Protocol Sharing
CDM Extractions
Code Sharing
Train / Test split

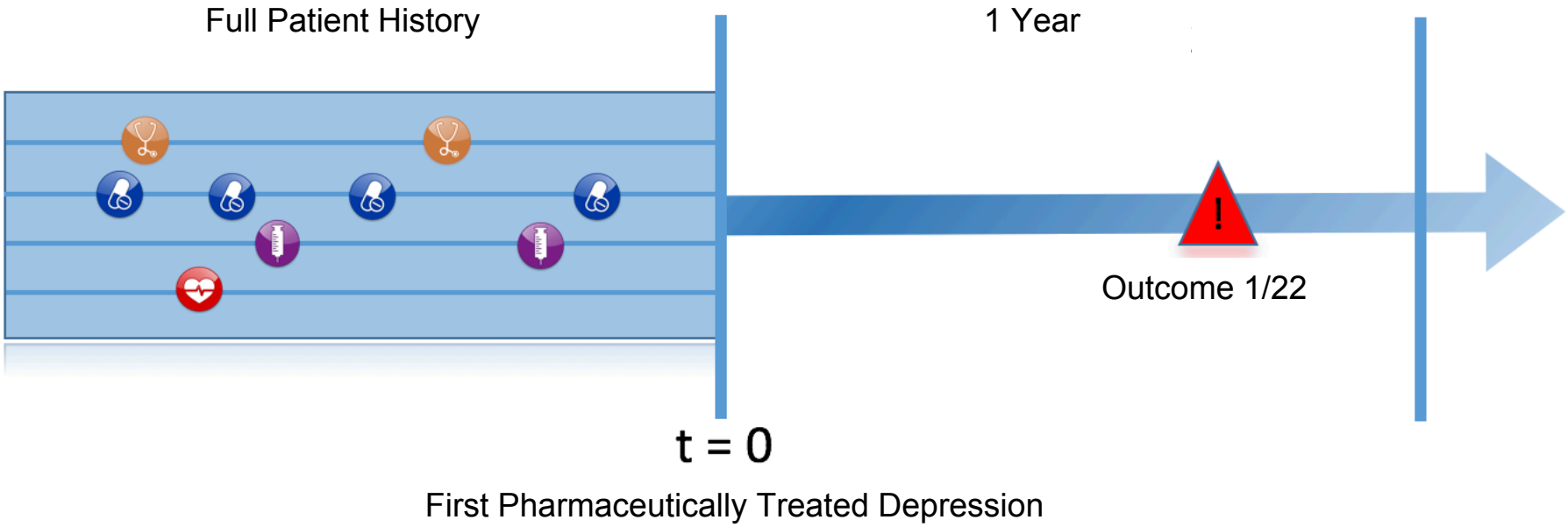
Standardization
Discrimination
Calibration
External Validation

Publications (TRIPOD)
Model sharing
Full transparency

Questions?

Part 3: Prediction in Patients with Pharmaceutically Treated Depression

Problem definition



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.

Target (T) 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

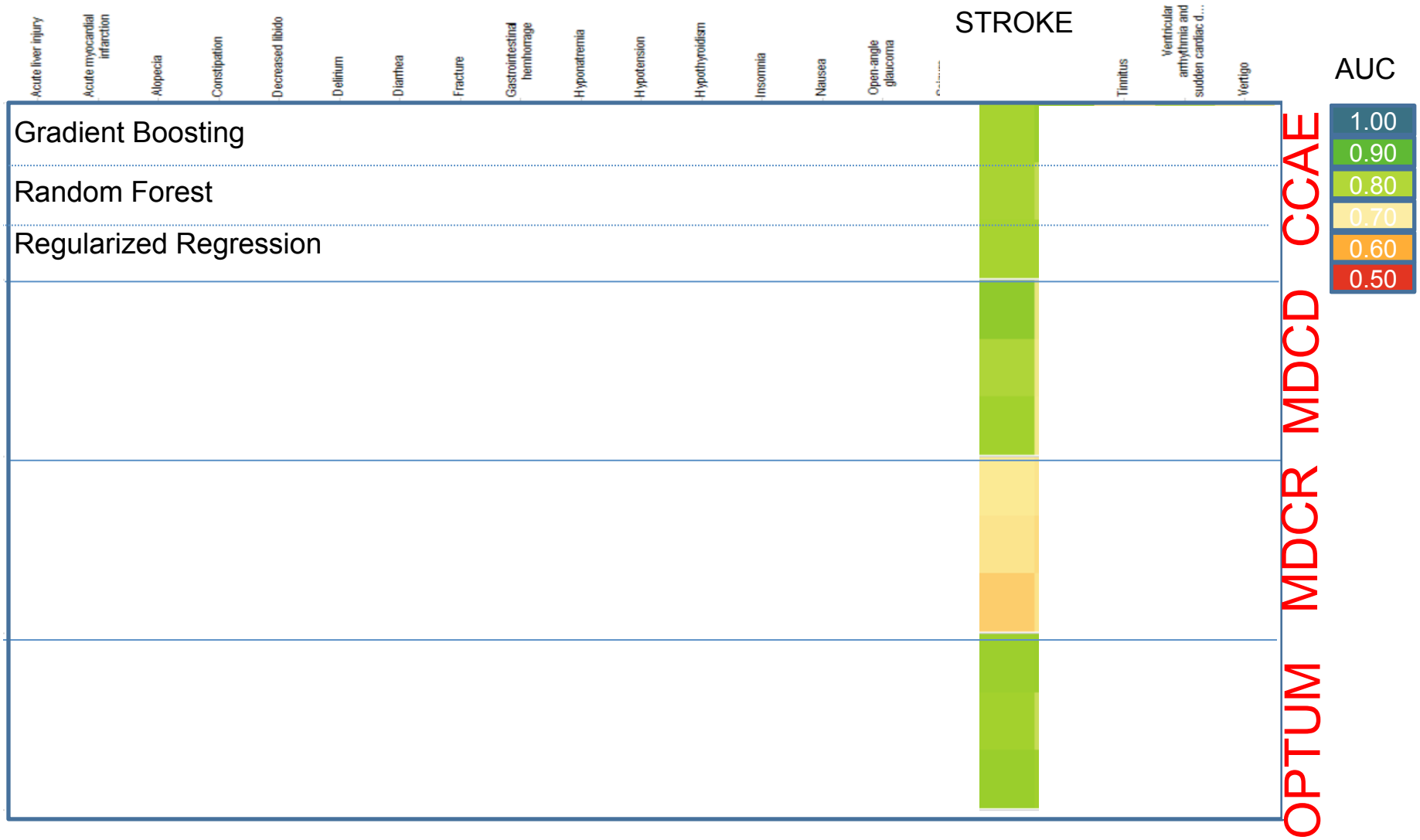
- Time split for training and testing
- Transportability for Stroke

Models

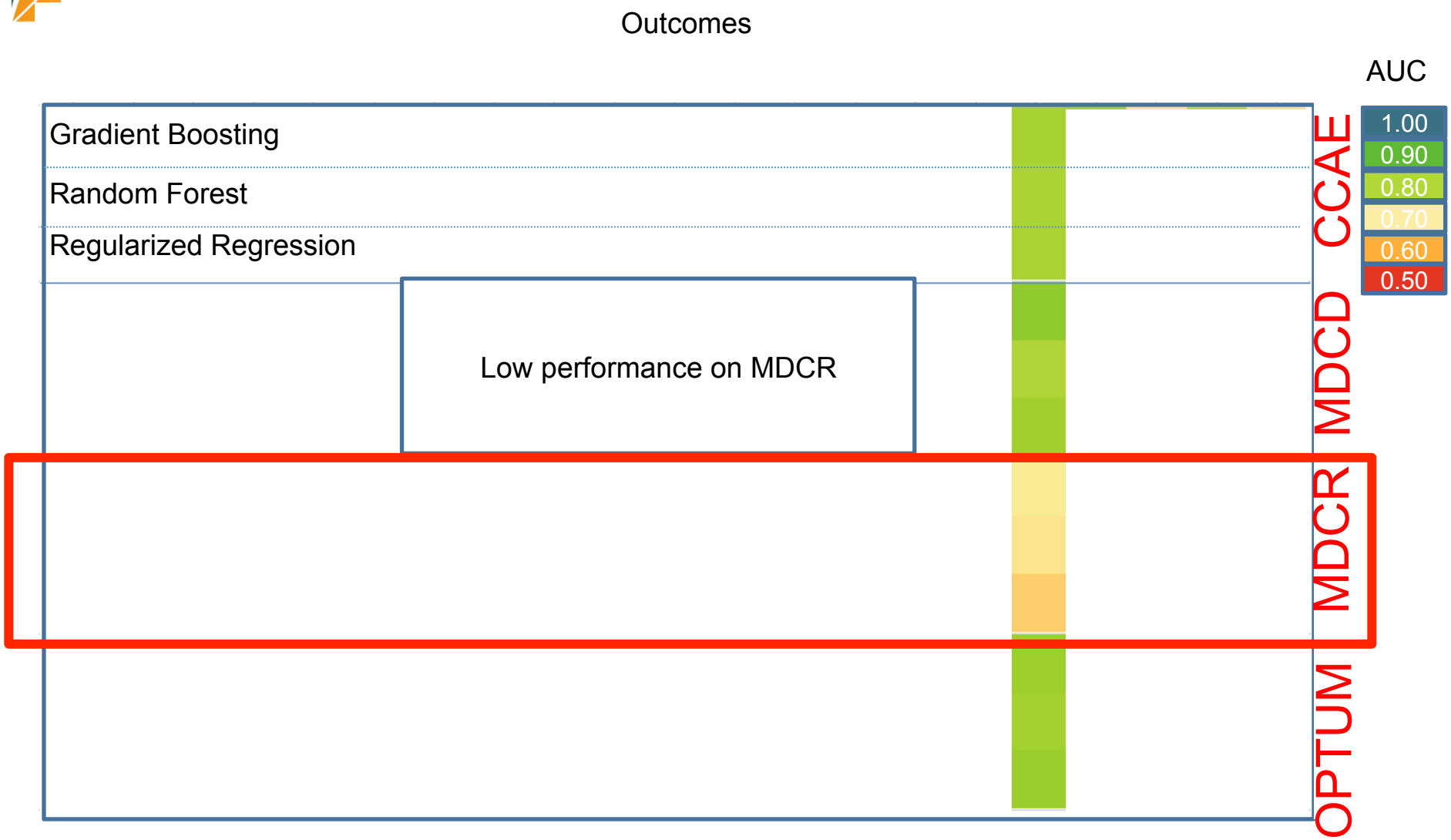
- Gradient Boosting
- Random Forest
- Regularized Regression

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

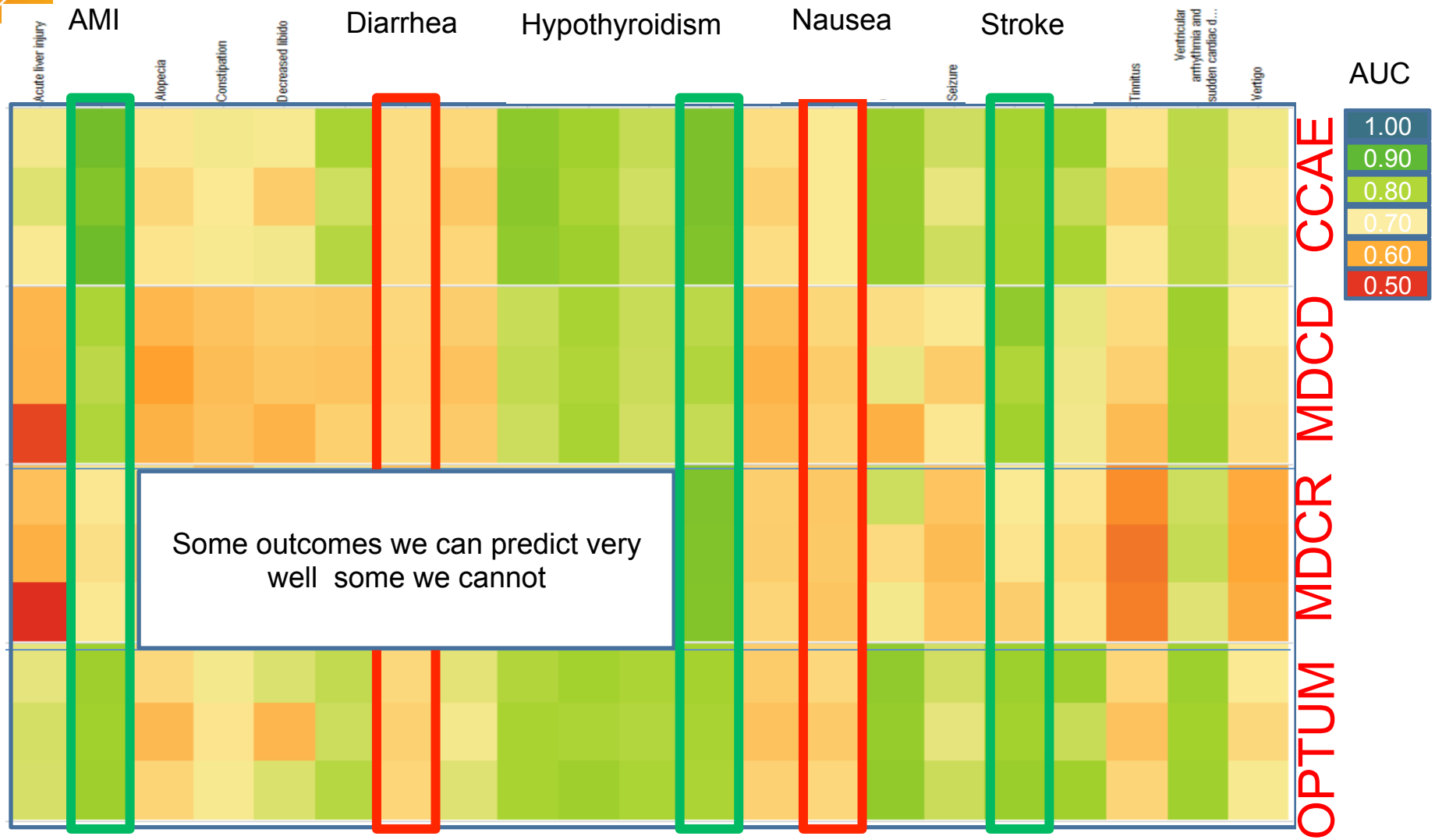
Model Discrimination Stroke



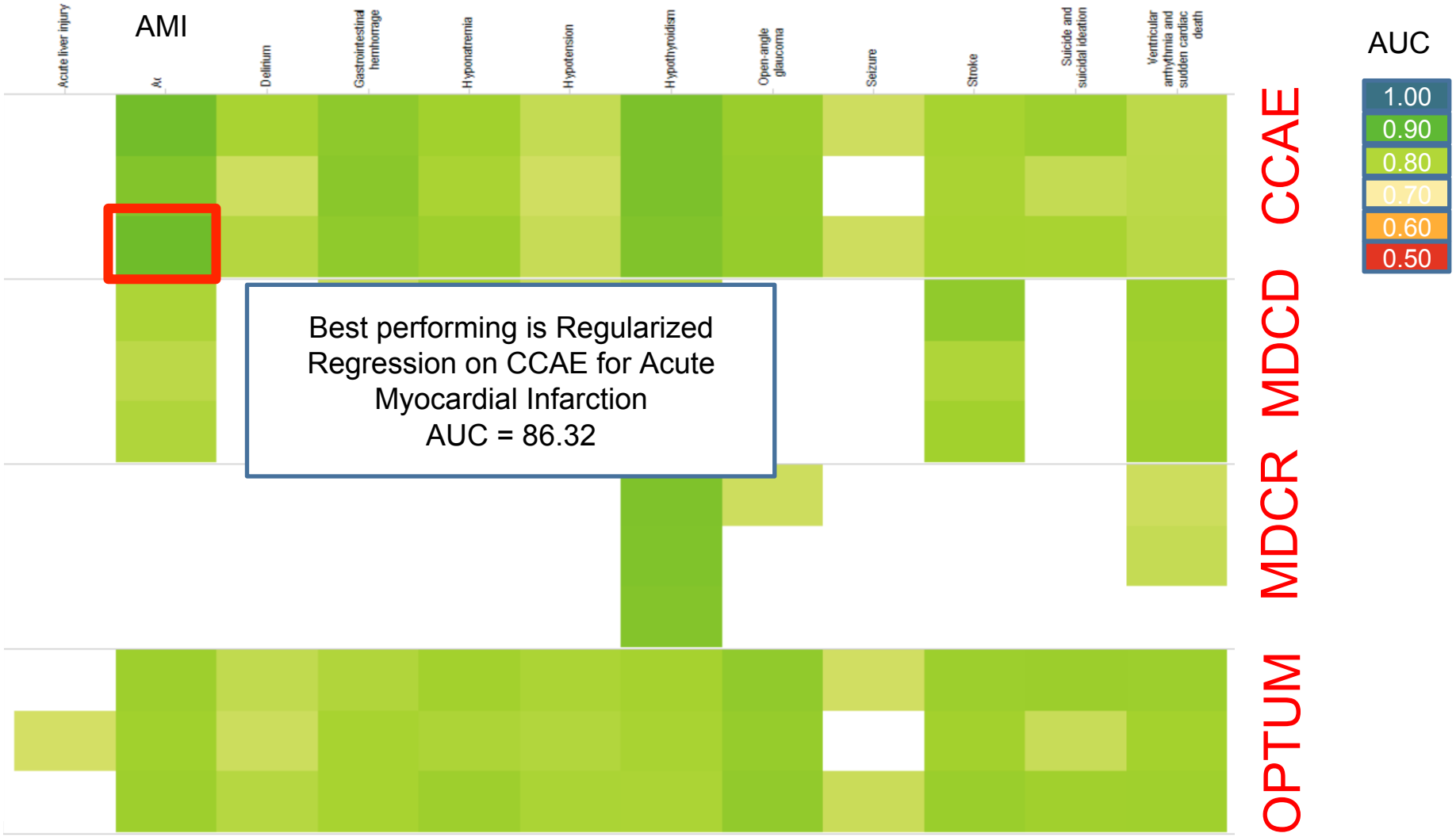
Model Discrimination



Model Discrimination

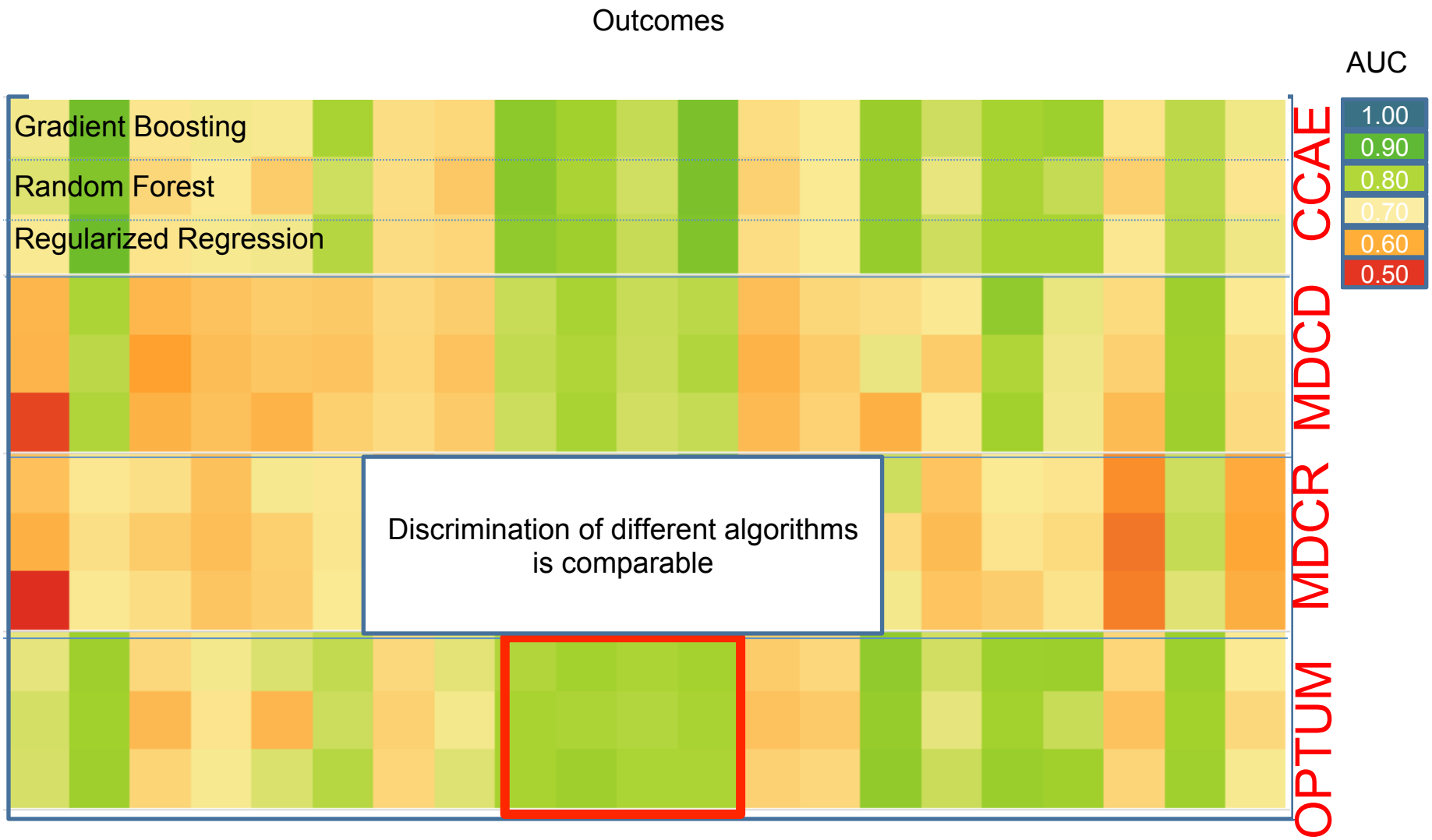


Outcomes with AUC > 0.75

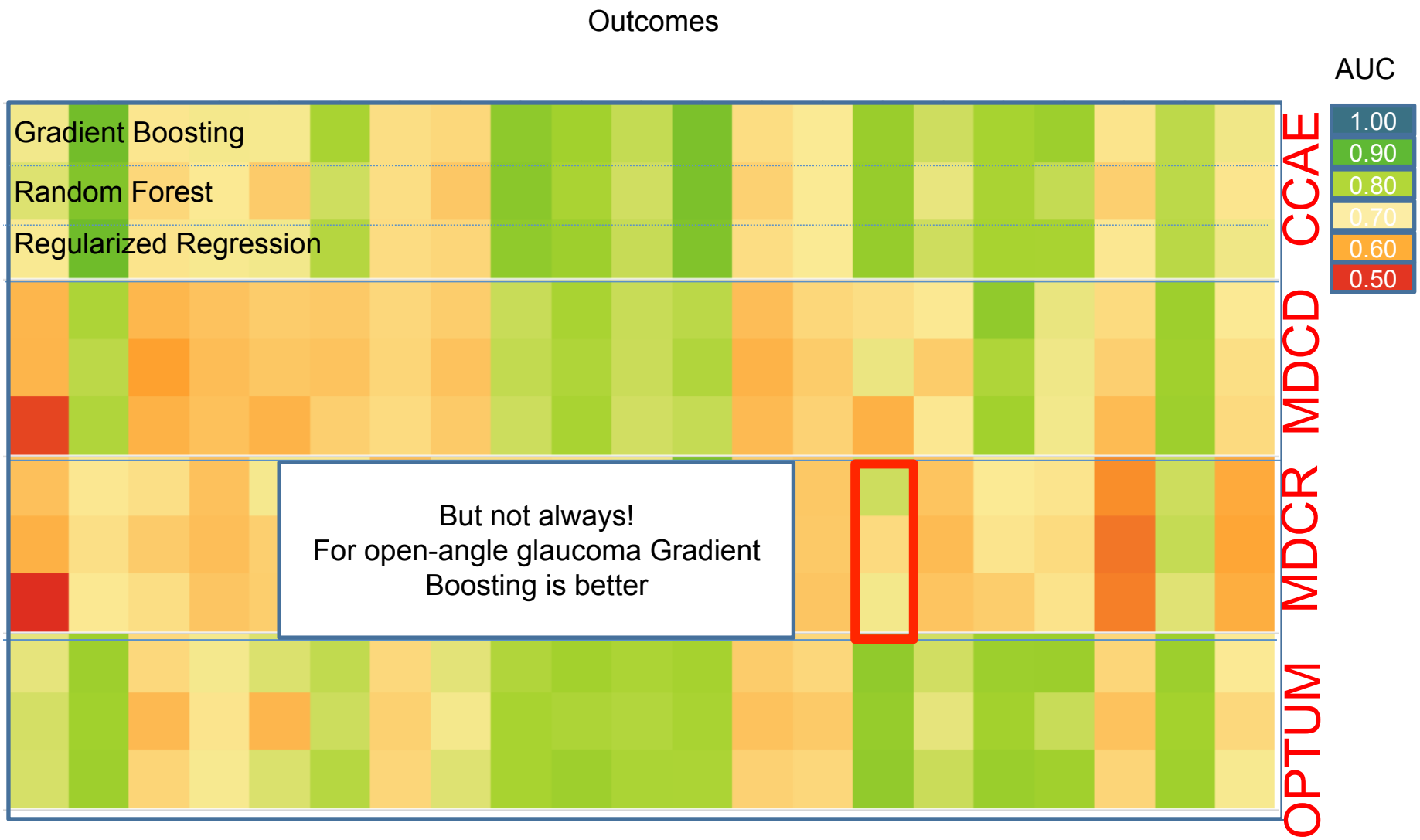


Best performing is Regularized Regression on CCAE for Acute Myocardial Infarction
AUC = 86.32

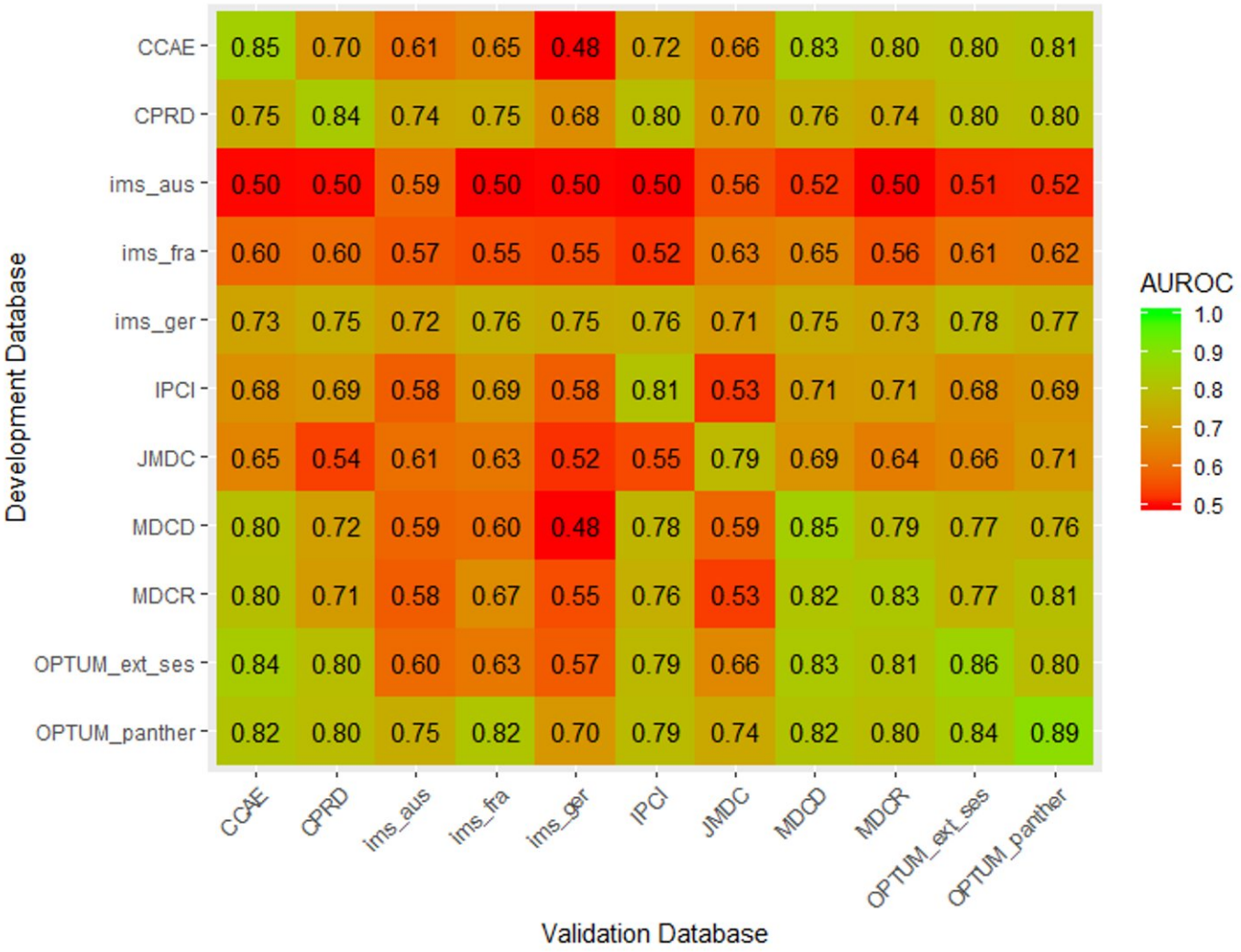
Model Discrimination



Model Discrimination



External Validation

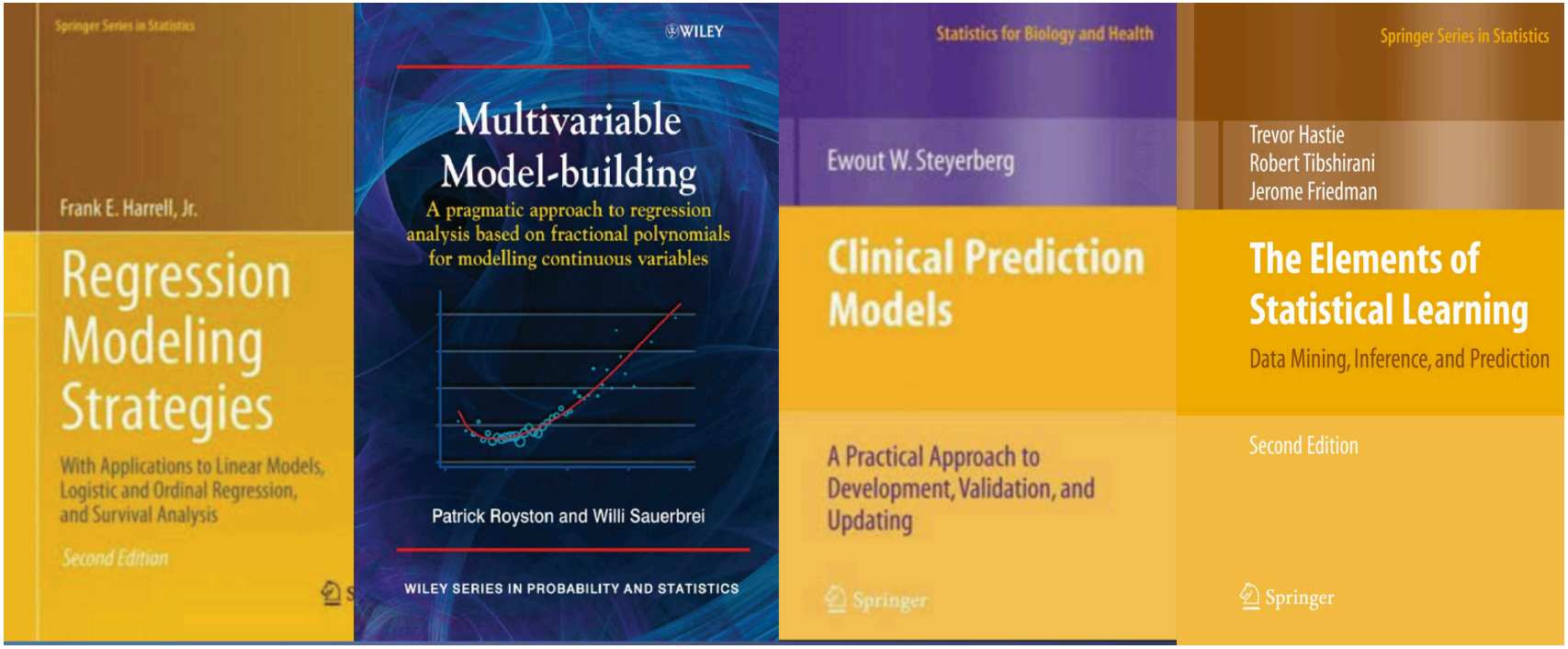


What did we achieve so far?

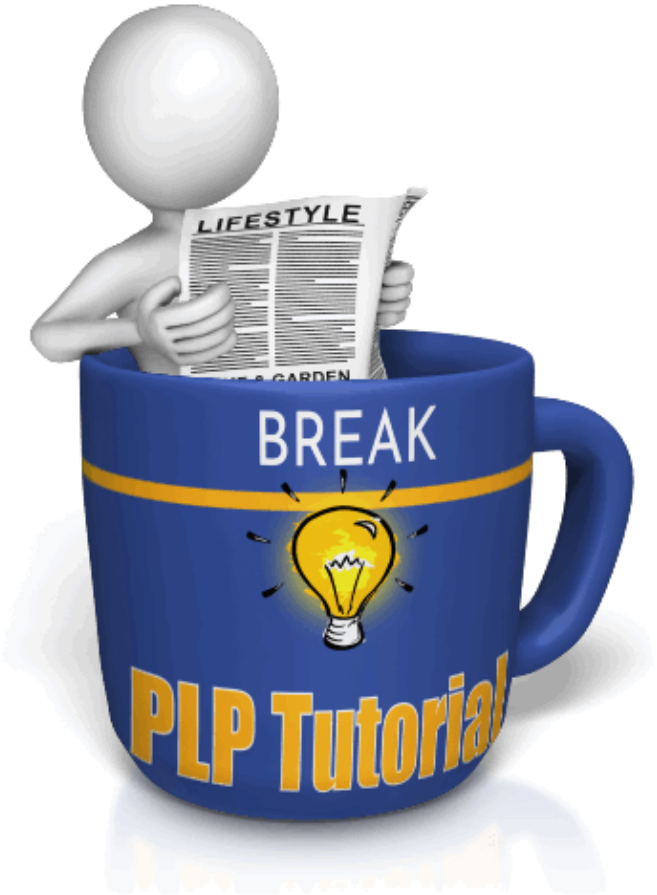
We showed it is feasible to develop large-scale predictive models for all databases converted to the OMOP CDM. This can now be done for any target cohort (T), outcome (O), and time at risk.

Further Reading if you got very interested!

- Phases of Clinical Prediction Modeling BMJ Series 2009
- Many good textbooks:



Let’s take a 15 min break

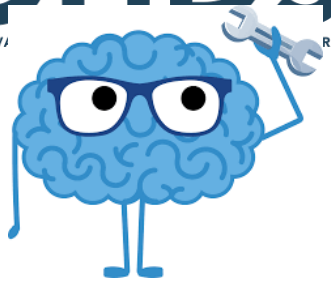


Today's Agenda

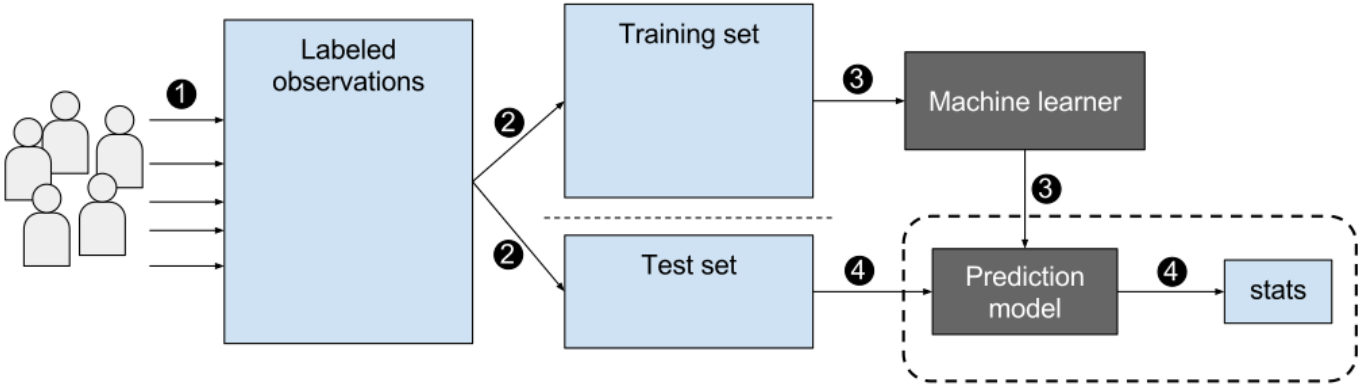
Time	Topic
8:45 – 9:00	Get settled, get laptops ready
9:00 – 10:00	Exercise: Selection of prediction problem
10:00 – 10:45	Presentation: What is Patient-Level Prediction
10:45 – 11:00	Break
11:00 – 11:45	Presentation: Learning the OHDSI Patient-Level Prediction Framework
11:45 – 12:30	Presentation: Overview of the TRIPOD Statement Exercise: Applying TRIPOD to CHADS2
12:30 – 13:15	Lunch
13:15 – 15:15	Guided tour through implementing patient-level prediction
15:15 – 15:30	Break
15:30 – 16:45	Exercise: Design and implement your own patient-level prediction
16:45 – 17:00	Lessons Learned and Feedback

Learning The PLP Framework

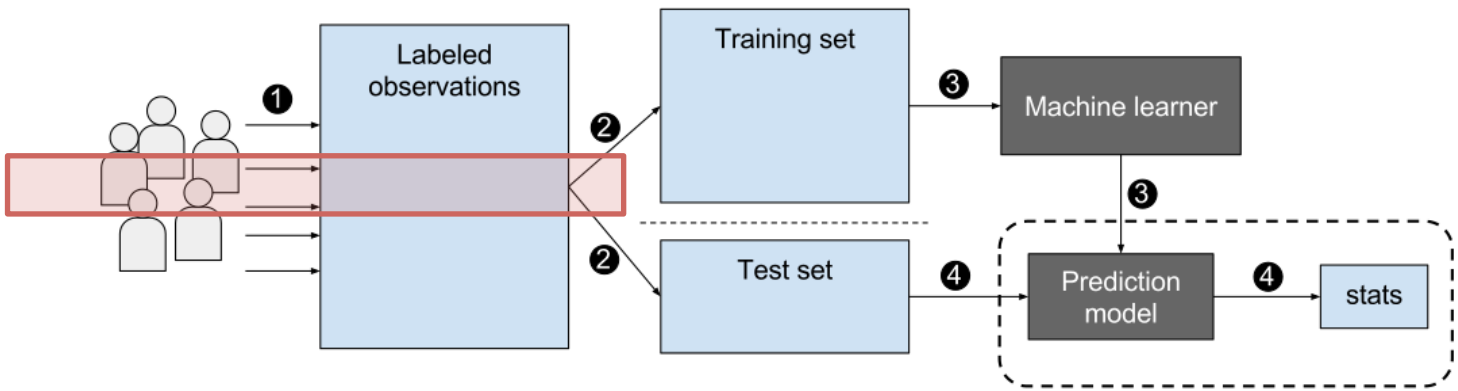
Understanding the components



Prediction Process



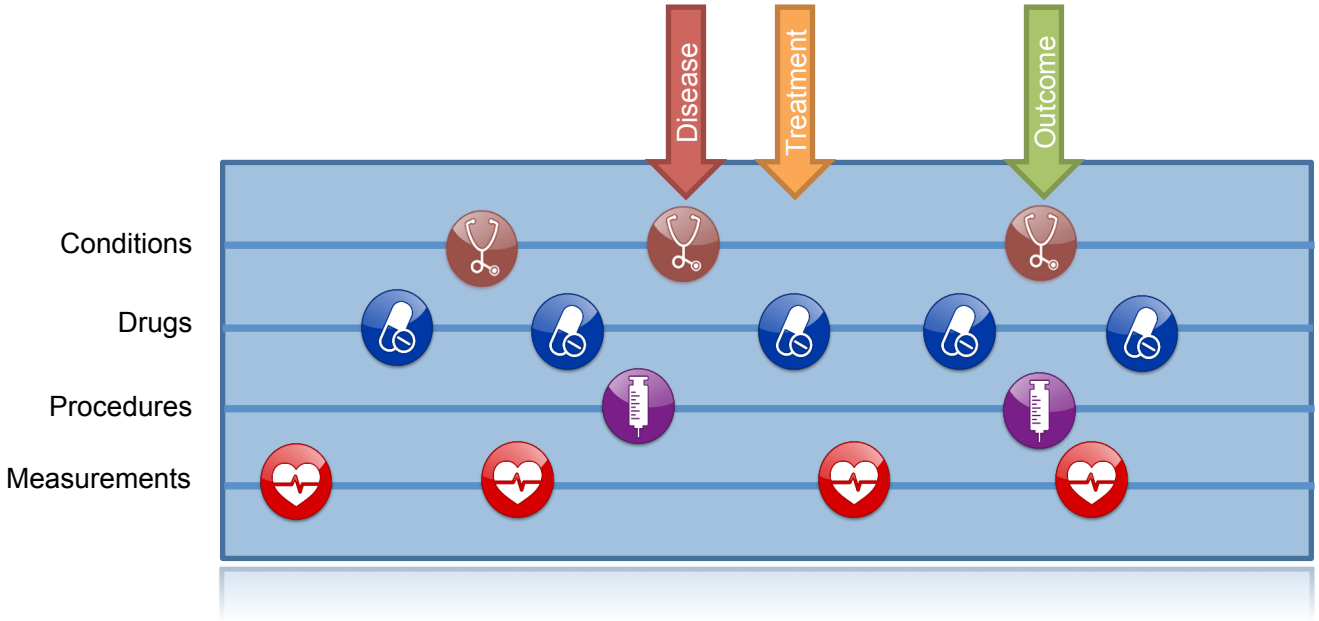
Prediction Process



What is our labeled data?

Our Data

- We have longitudinal data but we need labelled data for prediction…



Person ID	Concept ID	Date
1	343	2016-01-01
1	12045	2016-01-12
1	88466	2017-04-05
…		
1	0945	2019-01-23
2	343	2010-12-03
2	635636	2010-12-03
2	543	2010-12-05
…		

Defining Prediction Problem

- You need a well defined and clear prediction problem

- Considerations:
  - Is this clinically useful?
  - Is there a clear timepoint to apply the model?

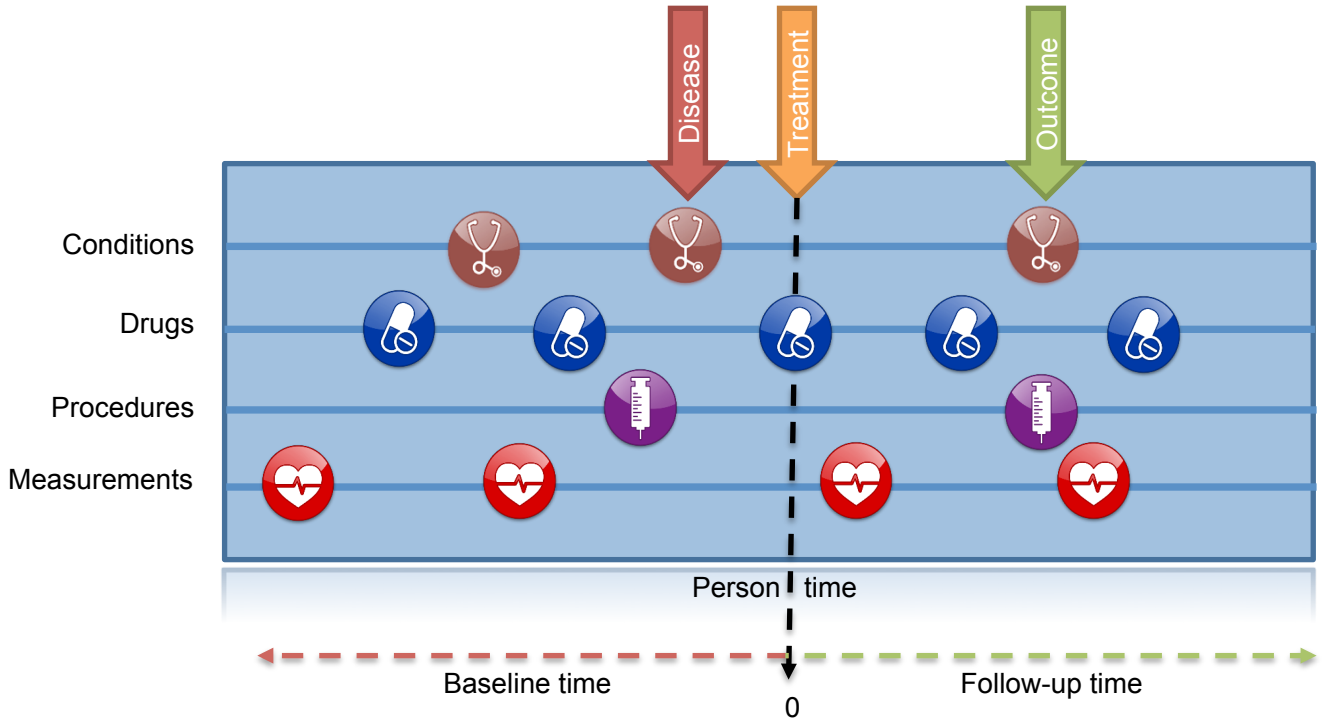
Prediction Question

In **<target population>** predict who develop **<outcome>** during **<time-at-risk>**

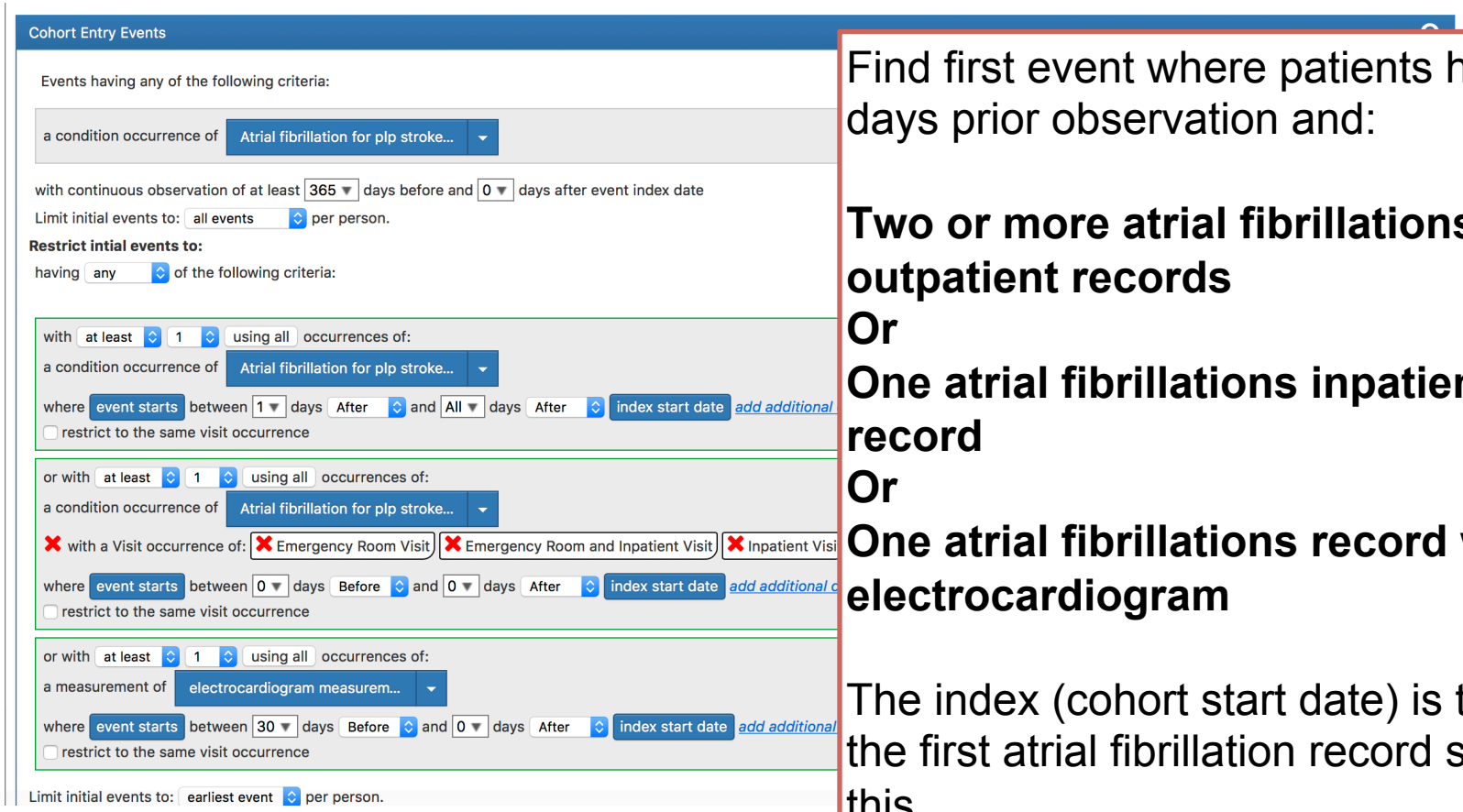
- **<target population> :** *The population of patients who you want to apply the model to, e.g., pregnant women, new users of drug X, those newly diagnosed with condition Y*
- **<outcome> :** *The thing you want to predict, e.g., death, stroke, depression*
- **<time-at-risk>** : *The period of time you want to predict the outcome occurring relative to the target population index date, e.g., 1 day until 365 days after index*

Extracting Labelled Data

The target cohort is key because we need an **index date** to pivot on… Data **before** index are used as **features** and data **after** index are used to see whether **outcome occurs during TAR**



Target Cohort Logic for Atrial Fibrilation



Find first event where patients have 365 days prior observation and:

Two or more atrial fibrillations outpatient records
Or
One atrial fibrillations inpatient/ER record
Or
One atrial fibrillations record with an electrocardiogram

The index (cohort start date) is the date of the first atrial fibrillation record satisfying this

Target Cohort Table for Atrial Fibrilation

A unique identifier for a patient

The cohort the patient belongs to (e.g., 1= atrial fibrillation)

The day a patient enters/exits the cohort – one of these is the index date (e.g., when they have atrial fibrillation)

Subject_id	Cohort_definition_id	Cohort_start_date	Cohort_end_date
3454102	1	2012-01-02	2012-01-01
105454	1	2012-08-12	2012-08-12
105459	1	2009-05-05	2009-05-05
4346356	1	2011-07-05	2011-07-05
342424	1	2010-01-01	2010-01-01
…	…	…	…

Outcome Cohort Logic for Ischemic Stroke

Cohort Entry Events

Events having any of the following criteria:

+ Add Initial Event

a condition occurrence of [LEGEND HTN] Ischemic stroke

with continuous observation of at least 0 days before and 0 days after event index date

Limit initial events to: all events per person.

Restrict intial events to:

having any of the following criteria:

with at least 1 using all occurrences of:

a visit occurrence of Inpatient or ER visit

where event starts between All days Before and 1 days After index start date

✖ and event ends between 0 days Before and All days After index start date

restrict to the same visit occurrence

Limit initial events to: all events per person.

Remove initial event restriction

Find all events of:

Ischemic stroke record within an inpatient or ER visit

The index (outcome start date) is the date of the inpatient ischemic stroke

Outcome Cohort For Ischemic Stroke

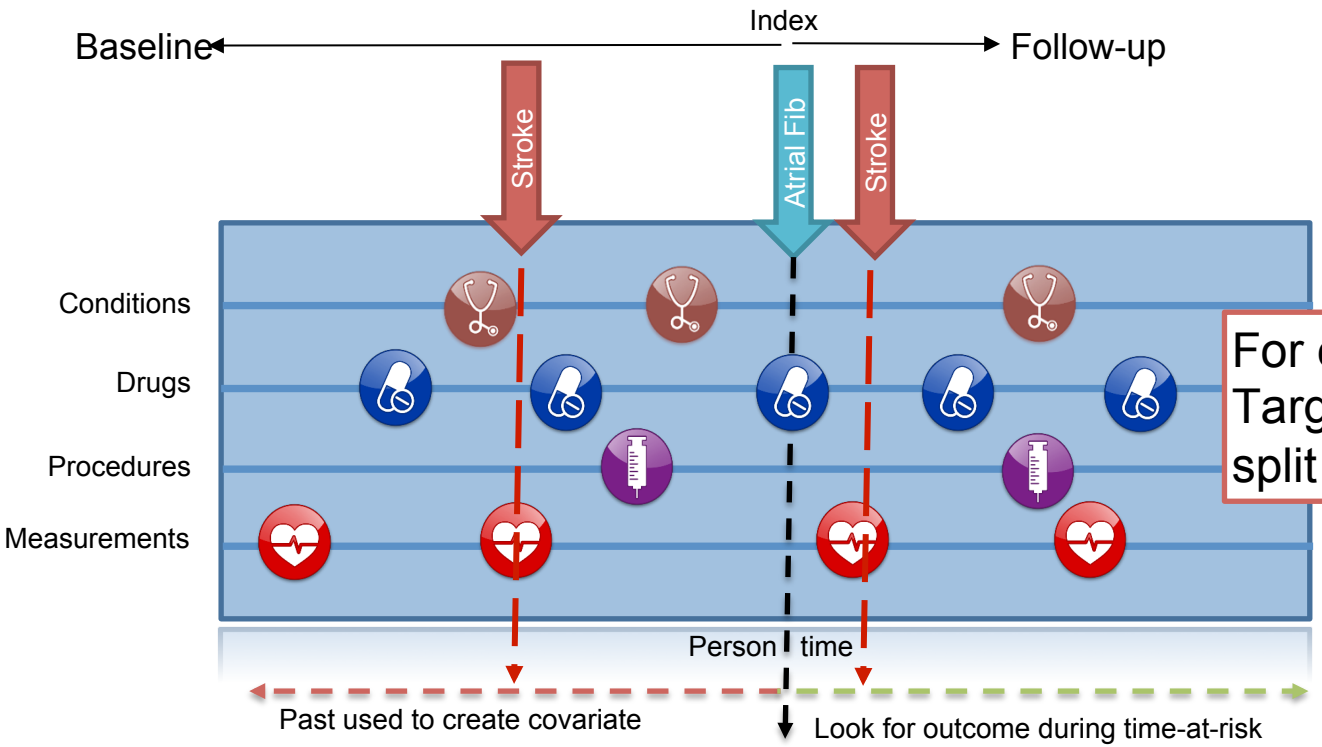
A unique identifier for a patient

The cohort the patient belongs to (e.g., 2= ischemic stroke)

The day a patient enters/exits the cohort – one of these is the index date (e.g., when they have stroke)

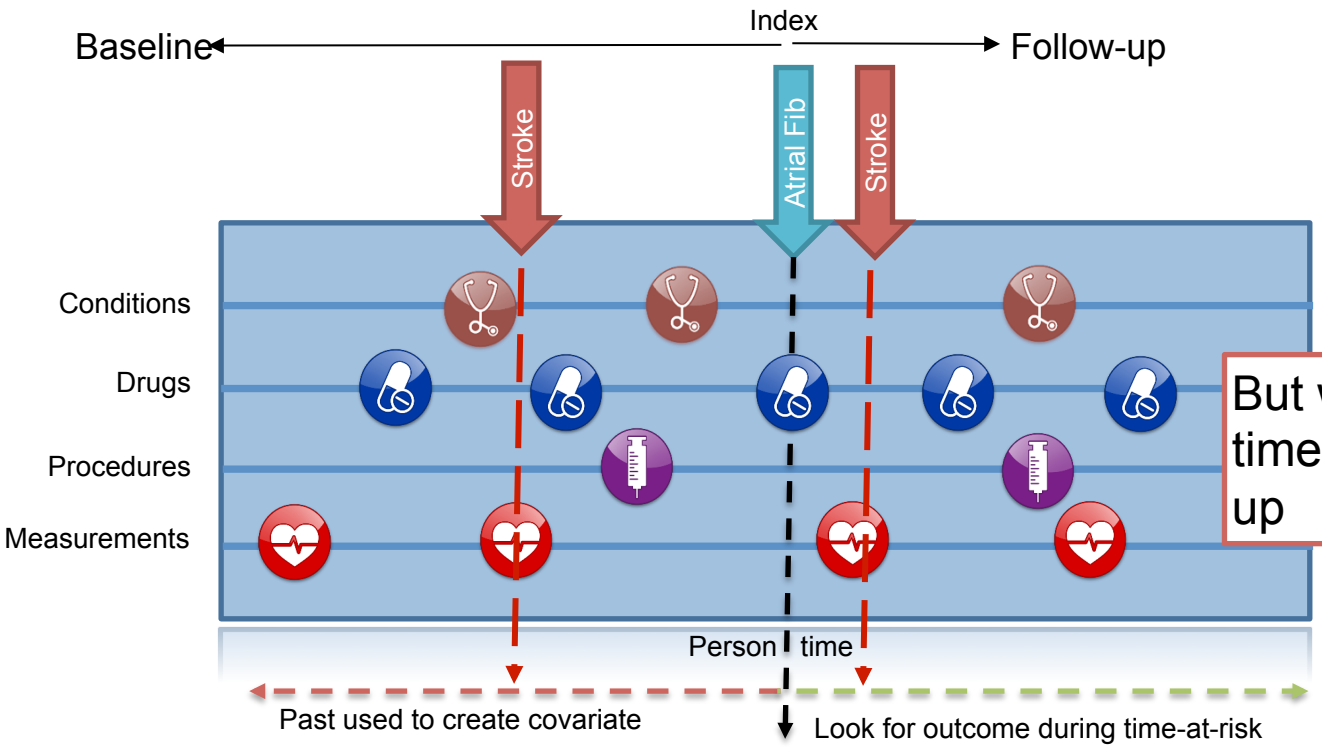
Subject_id	Cohort_definition_id	Cohort_start_date	Cohort_end_date
4346356	2	2010-09-12	2010-09-12
4346356	2	2011-08-01	2011-08-01
342424	2	2012-02-01	2012-02-01
1009833	2	2016-04-05	2016-04-05
…	…	…	…

Extracting Labelled Data



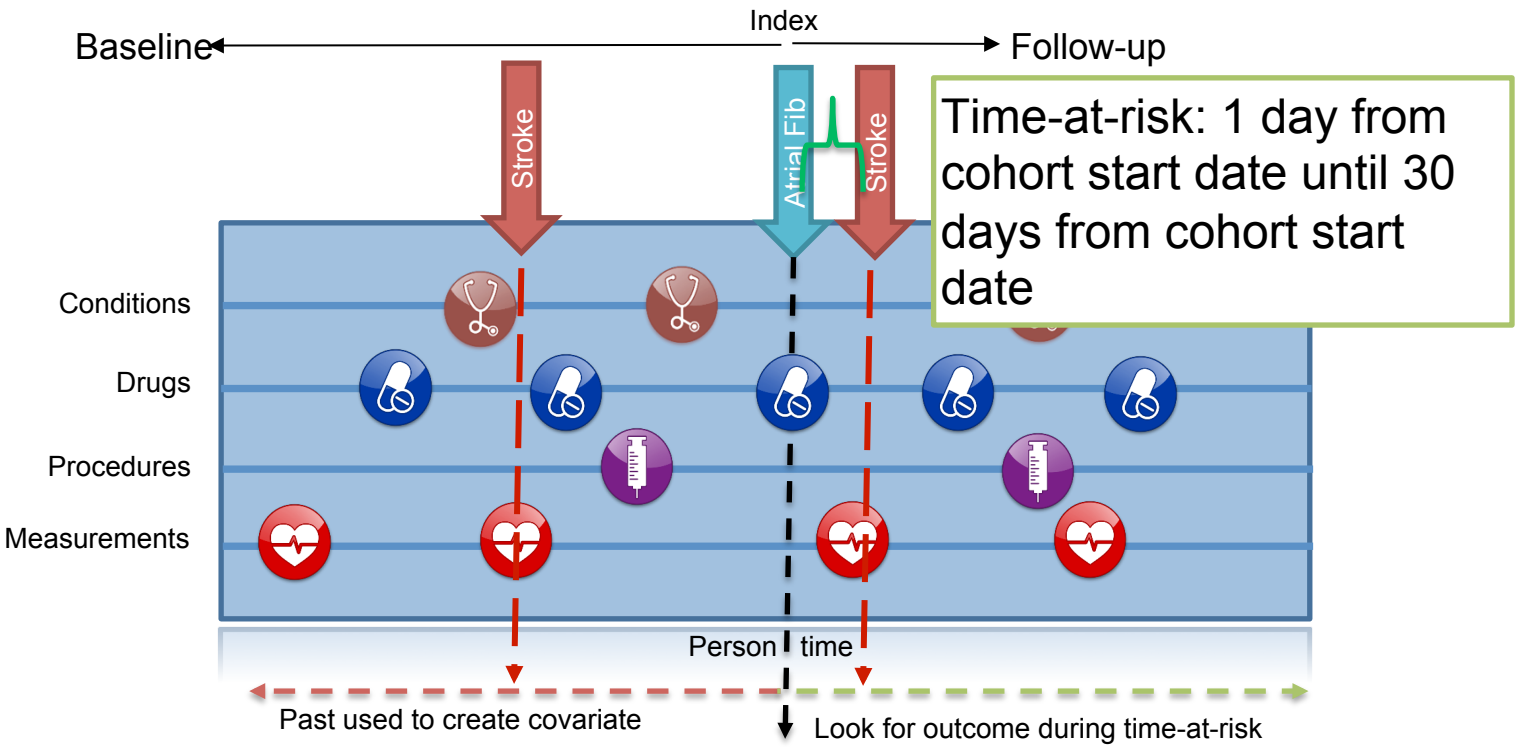
For each patient in the Target cohort we can now split their time at the index

Extracting Labelled Data

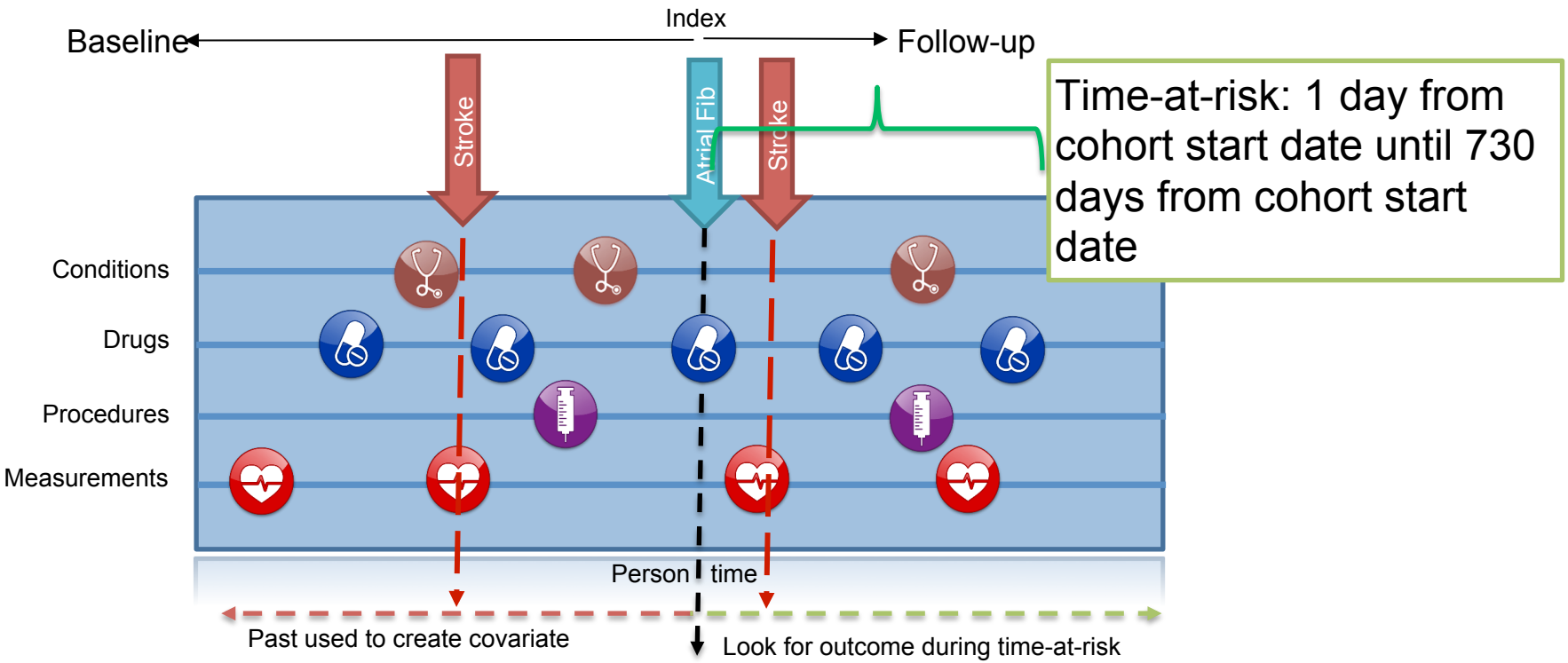


But we need to define the time-at-risk in the follow-up

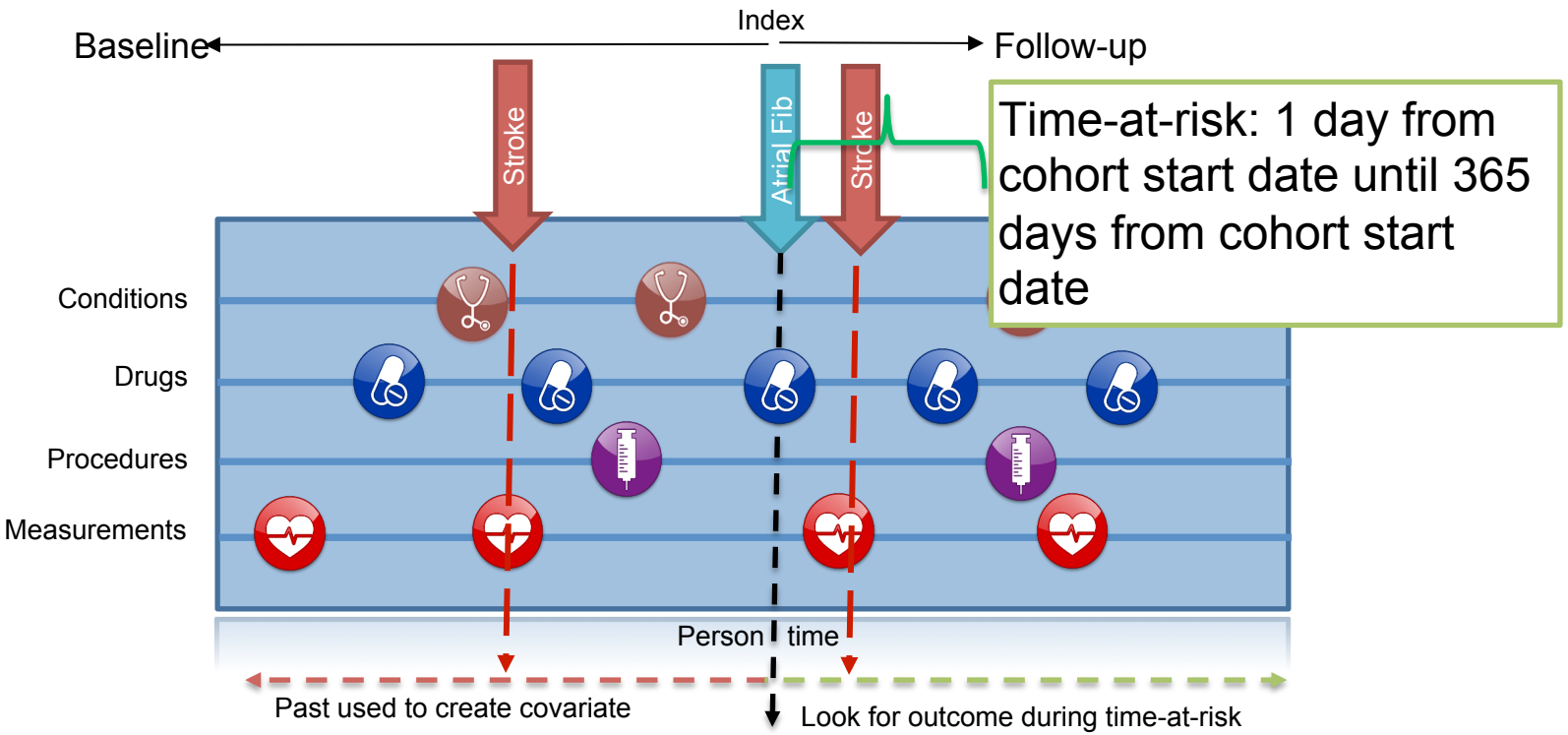
Extracting Labelled Data



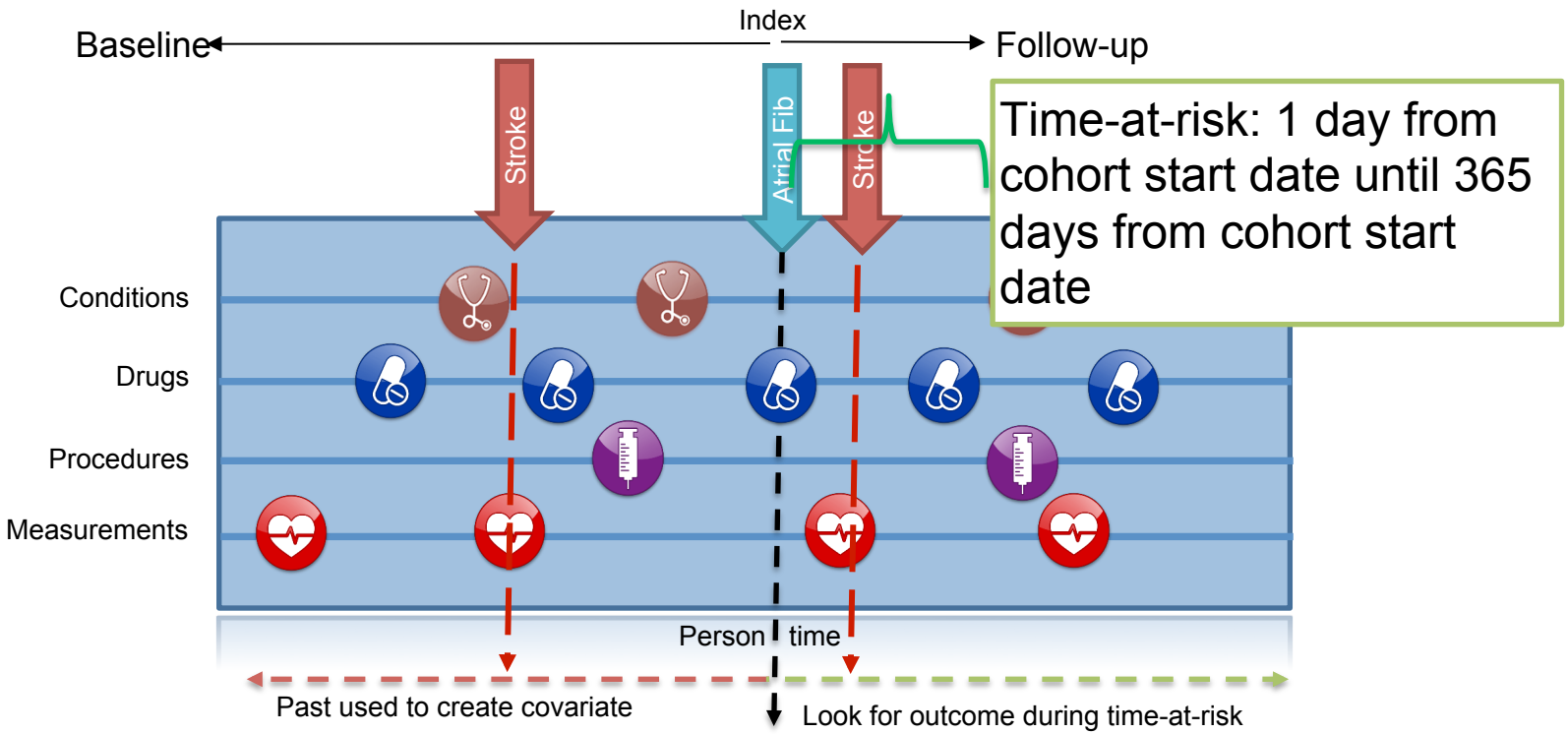
Extracting Labelled Data



Extracting Labelled Data



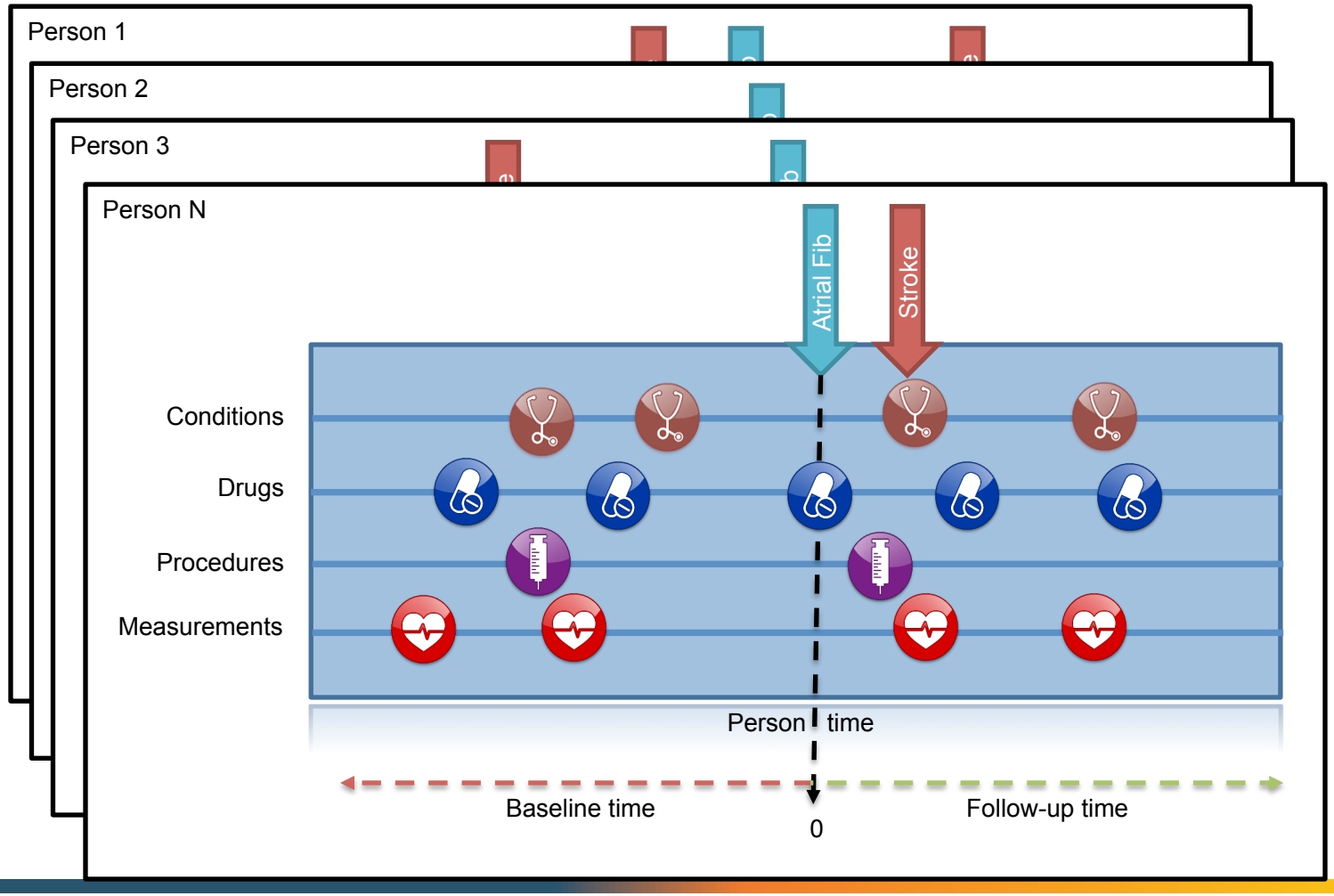
Extracting Labelled Data



Condition A	Condition B	…	Drug 1	…
1	0		0	

Outcome
1

We have this for many patients



Extracting Labelled Data

- Each person corresponds to a row

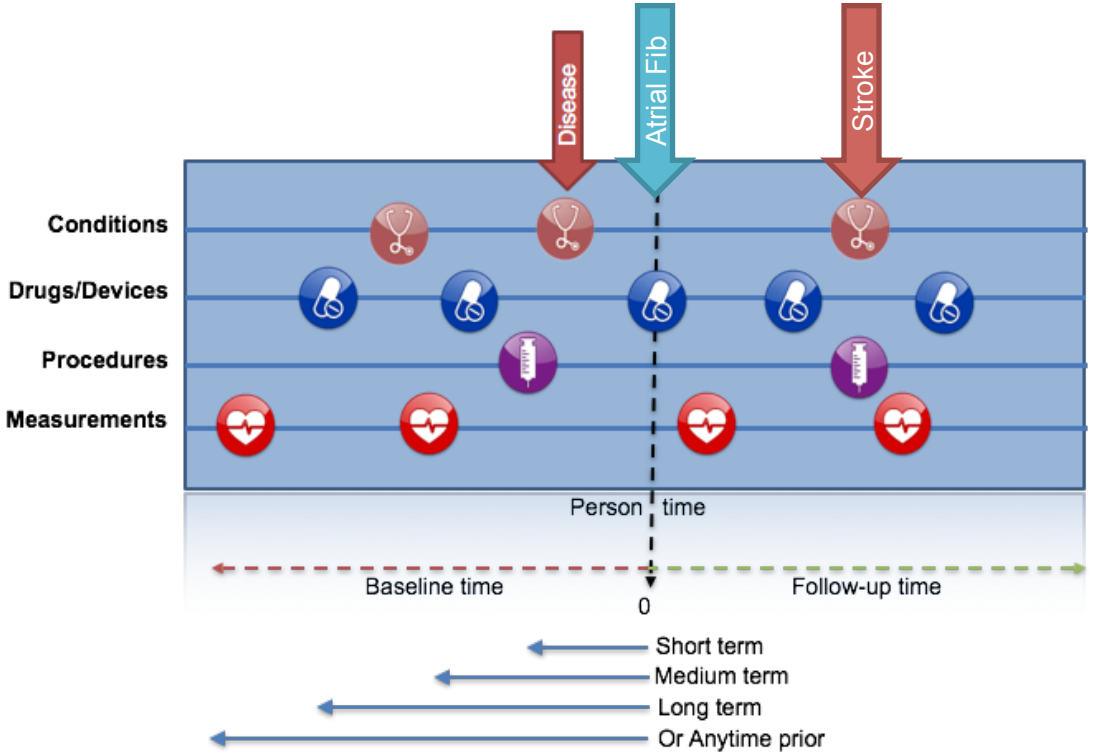
Labelled classification data

Subject_id	Cohort start date	Has outcome during TAR
3454102	2012-01-02	1 (Yes)
105454	2012-08-12	0 (No)
…		…

This gives us our labels for each subject!

Now Use Baseline to Construct Covariates

- We create standard features using records prior to the target cohort start date (e.g.,



Covariates

- Can pick three time periods and anytime prior to index (include index is an option)
- Binary indicator variables for conditions, drugs, procedures, measurements and observations
- Values for measurements
- Can use hierarchy to create binary indicators for a code and all children code (grouped covariates)
- Includes record type counts
- Includes some common risk scores
- Can add custom variables

Extracting Labelled Data

- We create the covariates using the baseline for each subject

Labelled classification data

Subject _id	Condition A	Condition B	…	Drug N	Has outcome during TAR
3454102	1	1	…	0	1 (Yes)
105454	1	0	…	1	0 (No)
…	…	…	…	…	…

This gives us our label data for each subject!

Population Settings

We have extra inclusion settings in the framework

- Do you want to remove people who have the outcome prior (i.e., predict new occurrence of outcome)?
- Do you want to only include each person in the target population once?
- Do you want a minimum prior observation time (i.e., only include subjects with 3 years or prior records)?
- How do you want to deal with people who are lost to follow-up?

Population Settings

We have extra inclusion settings in the framework

- Do you want to remove people who have the outcome prior (i.e., predict new occurrence of outcome)?
- Do you want to only include each person in the target population once?
- Do you want a minimum prior observation time (i.e., only include subjects with 3 years or prior records)?
- How do you want to deal with people who are lost to follow-up?

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Prior outcome
3454102	1	2012-01-02	0	
105454	1	2012-08-12	0	
1554	1	2009-05-05	0	
56566	1	2011-07-05	0	Yes (-35 days and -999 days)
4346356	1	2011-07-05	1	
342424	1	2010-01-01	1	Yes (-370 days)
…	…	…		

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Prior outcome
3454102	1	2012-01-02	0	
105454	1	2012-08-12	0	
1554	1	2009-05-05	0	
56566	1	2011-07-05	0	Yes (-35 days and -999 days)
4346356	1	2011-07-05	1	
342424	1	2010-01-01	1	Yes (-370 days)
…	…	…		

- Remove patients who have observed the outcome prior to cohort entry? **[YES]**
- How many days to look back from cohort entry for the outcome? **[99999]** days prior to cohort start

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Prior outcome
3454102	1	2012-01-02	0	
105454	1	2012-08-12	0	
1554	1	2009-05-05	0	
56566	1	2011-07-05	0	Yes (-35 days and -999 days)
4346356	1	2011-07-05	1	
342424	1	2010-01-01	1	Yes (-370 days)
…	…	…		

- Remove patients who have observed the outcome prior to cohort entry? **[YES]**
- How many days to look back from cohort entry for the outcome? **[365]** days prior to cohort start

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Prior outcome
3454102	1	2012-01-02	0	
105454	1	2012-08-12	0	
1554	1	2009-05-05	0	
56566	1	2011-07-05	0	Yes (-35 days and -999 days)
4346356	1	2011-07-05	1	
342424	1	2010-01-01	1	Yes (-370 days)
…	…	…		

- Remove patients who have observed the outcome prior to cohort entry? **[No]**

Population Settings

We have extra inclusion settings in the framework

- Do you want to remove people who have the outcome prior (i.e., predict new occurrence of outcome)?
- Do you want to only include each person in the target population once?
- Do you want a minimum prior observation time (i.e., only include subjects with 3 years or prior records)?
- How do you want to deal with people who are lost to follow-up?

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR
3454102	1	2012-01-02	0
105454	*1*	*2012-08-12*	*0*
105454	*1*	*2013-10-04*	*0*
1554	1	2009-05-05	0
56566	1	2011-07-05	0
4346356	1	2011-07-05	1
342424	1	2010-01-01	1
…	…	…	

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR
3454102	1	2012-01-02	0
105454	*1*	*2012-08-12*	*0*
~~*105454*~~	~~*1*~~	~~*2013-10-01*~~	~~*0*~~
1554	1	2009-05-05	0
56566	1	2011-07-05	0
4346356	1	2011-07-05	1
342424	1	2010-01-01	1
…	…	…	

Should only the first exposure per subject be included? **[YES]**

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR
3454102	1	2012-01-02	0
105454	*1*	*2012-08-12*	*0*
105454	*1*	*2013-10-04*	*0*
1554	1	2009-05-05	0
56566	1	2011-07-05	0
4346356	1	2011-07-05	1
342424	1	2010-01-01	1
…	…	…	

Should only the first exposure per subject be included? **[No]**

Population Settings

We have extra inclusion settings in the framework

- Do you want to remove people who have the outcome prior (i.e., predict new occurrence of outcome)?
- Do you want to only include each person in the target population once?
- Do you want a minimum prior observation time (i.e., only include subjects with 3 years or prior records)?
- How do you want to deal with people who are lost to follow-up?

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Prior observation
3454102	1	2012-01-02	0	366
105454	1	2012-08-12	0	2009
1554	1	2009-05-05	0	1098
56566	1	2011-07-05	0	365
4346356	1	2011-07-05	1	4056
342424	1	2010-01-01	1	588
…	…	…		

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Prior observation
~~3454102~~	~~1~~	~~2012-01-02~~	~~0~~	~~366~~
105454	1	2012-08-12	0	2009
1554	1	2009-05-05	0	1098
~~56566~~	~~1~~	~~2011-07-05~~	~~0~~	~~365~~
4346356	1	2011-07-05	1	4056
~~342424~~	~~1~~	~~2010-01-01~~	~~1~~	~~588~~
…	…	…		

Minimum lookback period applied to target cohort: **[730]**

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Prior observation
~~3454102~~	~~1~~	~~2012-01-02~~	~~0~~	~~366~~
105454	1	2012-08-12	0	2009
~~1554~~	~~1~~	~~2009-05-05~~	~~0~~	~~1098~~
~~56566~~	~~1~~	~~2011-07-05~~	~~0~~	~~365~~
4346356	1	2011-07-05	1	4056
~~342424~~	~~1~~	~~2010-01-01~~	~~1~~	~~366~~
…	…	…		

Minimum lookback period applied to target cohort: **[1200]**

Imaging your cohort looks like this:

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Prior observation
3454102	1	2012-01-02	0	366
105454	1	2012-08-12	0	2009
1554	1	2009-05-05	0	1098
56566	1	2011-07-05	0	365
4346356	1	2011-07-05	1	4056
342424	1	2010-01-01	1	588
…	…	…		

Minimum lookback period applied to target cohort: **[365]**

Population Settings

We have extra inclusion settings in the framework

- Do you want to remove people who have the outcome prior (i.e., predict new occurrence of outcome)?
- Do you want to only include each person in the target population once?
- Do you want a minimum prior observation time (i.e., only include subjects with 3 years or prior records)?
- How do you want to deal with people who are lost to follow-up?

Imaging your cohort looks like this:

TAR (time-at-risk) is 1 day to 365 days after cohort start date

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Follow-up observation
3454102	1	2012-01-02	0	50
105454	1	2012-08-12	0	1082
1554	1	2009-05-05	0	366
56566	1	2011-07-05	0	480
4346356	1	2011-07-05	1	40
342424	1	2010-01-01	1	500
…	…	…		

Imaging your cohort looks like this:

TAR (time-at-risk) is 1 day to 365 days after cohort start date

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Follow-up observation
~~3454102~~	~~1~~	~~2012-01-02~~	~~0~~	~~30~~
105454	1	2012-08-12	0	1082
1554	1	2009-05-05	0	366
56566	1	2011-07-05	0	480
~~4346356~~	~~1~~	~~2011-07-05~~	~~1~~	~~40~~
342424	1	2010-01-01	1	500
…	…	…		

- Should subjects without time at risk be removed? **[YES]**
- Minimum time at risk: **[364]** days
- Include people with outcomes who are not observed for the whole at risk period? **[NO]**

Imaging your cohort looks like this:

TAR (time-at-risk) is 1 day to 365 days after cohort start date

Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Follow-up observation
3454102	1	2012-01-02	0	30
105454	1	2012-08-12	0	1082
1554	1	2009-05-05	0	366
56566	1	2011-07-05	0	480
4346356	1	2011-07-05	1	40
342424	1	2010-01-01	1	500
…	…	…		

- Should subjects without time at risk be removed? **[YES]**
- Minimum time at risk: **[364]** days
- Include people with outcomes who are not observed for the whole at risk period? **[YES]**

Imaging your cohort looks like this:

TAR (time-at-risk) is 1 day to 365 days after cohort start date

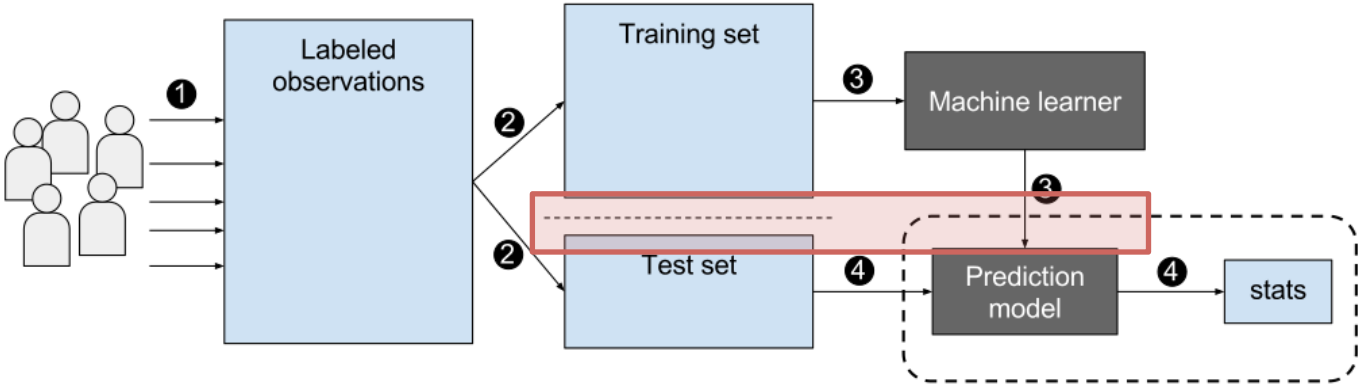
Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Follow-up observation
3454102	1	2012-01-02	0	50
105454	1	2012-08-12	0	1082
1554	1	2009-05-05	0	366
56566	1	2011-07-05	0	480
4346356	1	2011-07-05	1	40
342424	1	2010-01-01	1	500
…	…	…		

- Should subjects without time at risk be removed? **[No]**
- Minimum time at risk: **[1]** days
- Include people with outcomes who are not observed for the whole at risk period? **[No]**

Summary:

- Need to define prediction problem
- Need to define the target population and outcome cohorts
- Need to specify covariate settings
- Need to specify population settings – this modifies target population and creates labels

Prediction Process



Model Development Settings

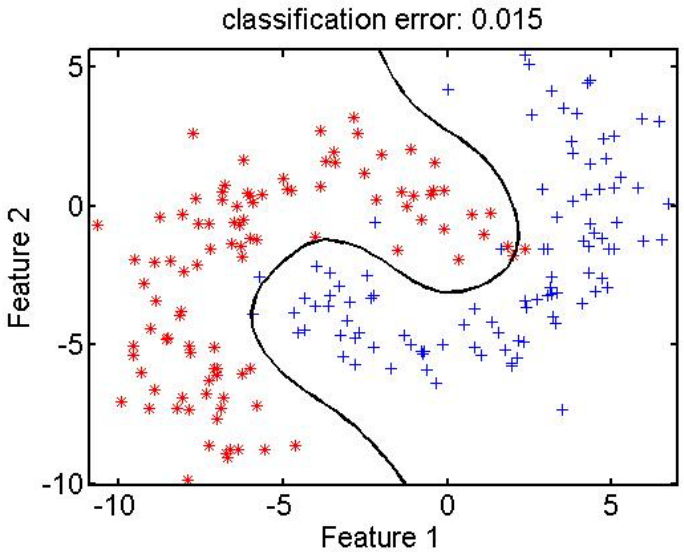
Train/Test Data Settings

- Train/Test %
  - e.g., 75/25
- Split Seed
  - e.g., 1
- Split type
  - e.g., time or person

Patient	Cohort start date	Person Split	Time Split
Patient 1	July 2016	Test	Test
Patient 2	Jan 1999	Train	Train
Patient 3	Jan 2001	Test	Train
Patient 4	June 2001	Train	Train
Patient 5	Feb 2016	Train	Test
Patient 6	Feb 2014	Train	Test
Patient 7	Nov 2003	Train	Train
Patient 8	Sept 2002	Train	Train
Patient 9	April 1998	Train	Train
Patient 10	April 2005	Test	Train
Patient 11	Dec 2008	Train	Train
Patient 12	March 2012	Train	Test
Patient 13	May 2010	Train	Train
Patient 14	Aug 2009	Test	Train
Patient 15	Aug 2009	Train	Train
Patient 16	Oct 2001	Train	Train

Training Classifier

- Learns to map covariates to class
- Effectively about learning a decision boundary that partitions the two classes
- Different classifiers lead to different decisions boundaries



Training Settings

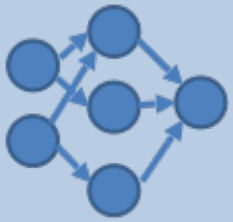
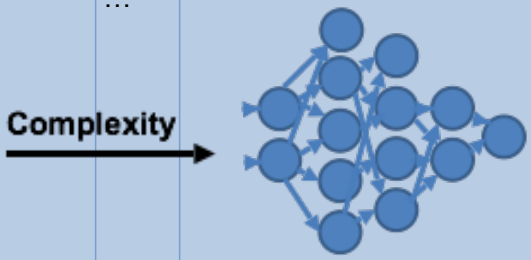
- Select the machine learning models that will be trained
- Define the hyper-parameter search strategy

Library of classifiers built in

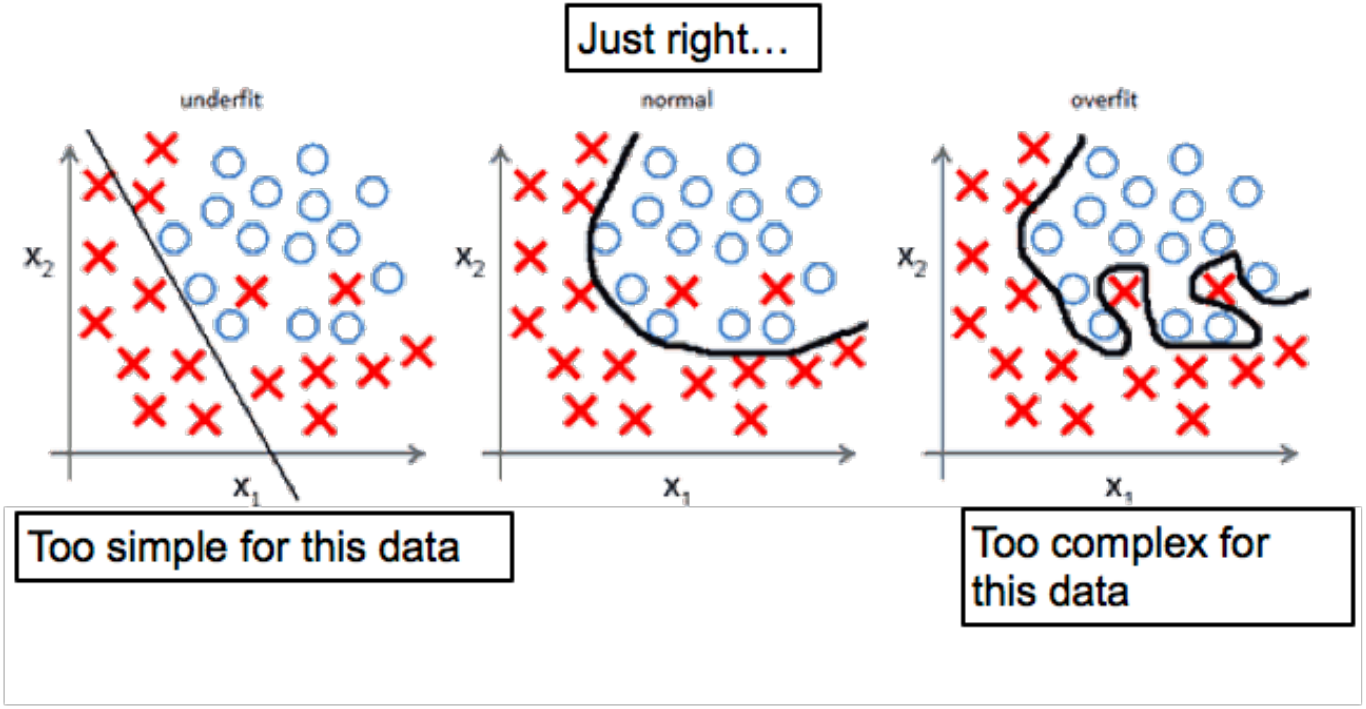
	Model
	Lasso Logistic Regression
	Gradient Boosting Machine
	Random Forest
	Adaboost
	Decision Tree
	Neural Network/Deep Learning
	K-nearest neighbours
	Naïve Bayes
?	Your custom model

What are hyper-parameters?

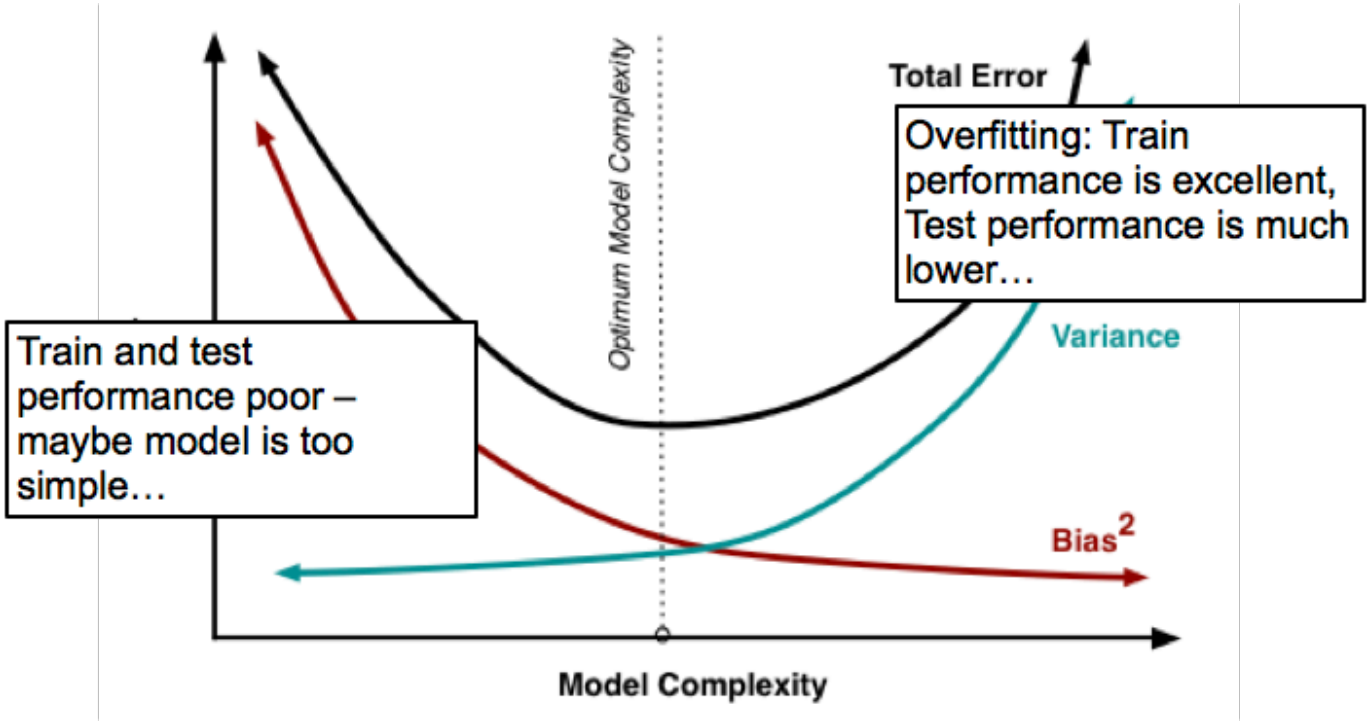
- They control the complexity of a model
- E.g., if we wanted to fit a neural network the topology of the network defines the complexity of the model (few layers and a small number of nodes = more simple)

Simple Model	...	Complex Model
	... 	
High bias/low variance (unlikely to overfit but may not be able to model complexities…)	...	High variance/low bias (prone to overfitting…)

Over vs Under fitting



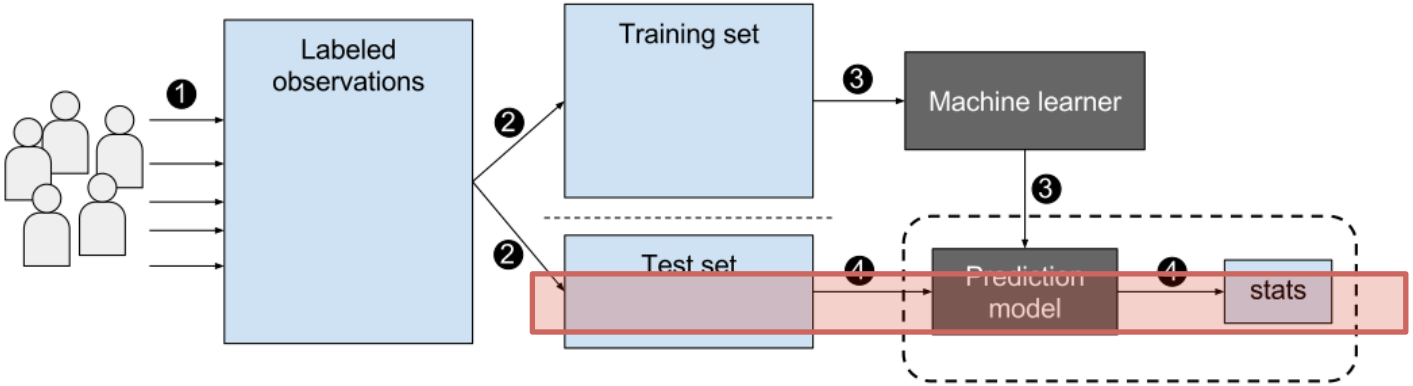
Over vs Under fitting



Summary:

- We suggest trying multiple classifiers
- We have a large library of classifiers but you can also add custom ones
- We have default hyper-parameter grid searches but you can expand this

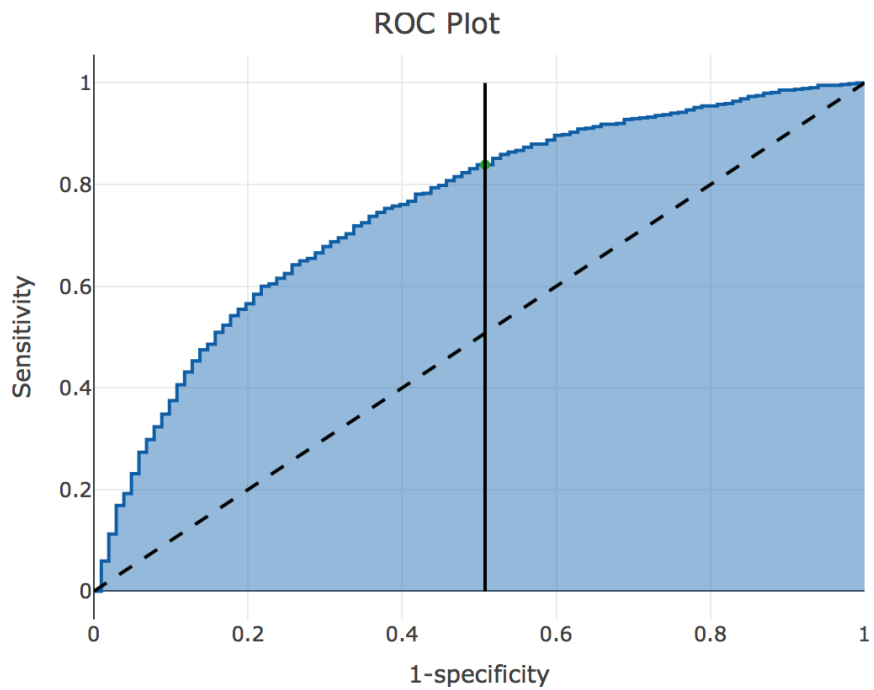
Prediction Process



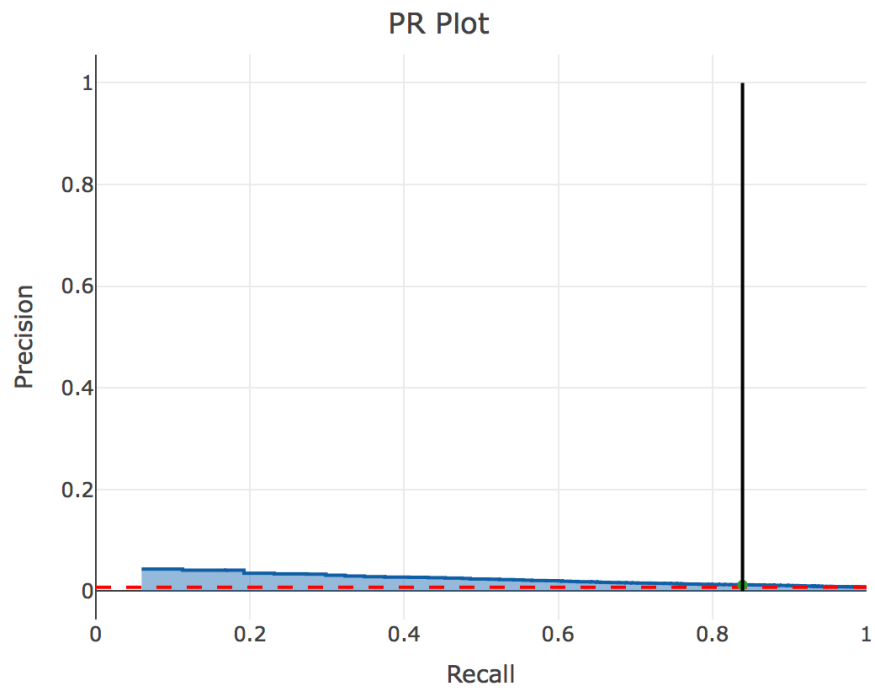
Model Evaluation

Metrics/Plots

ROC Plot

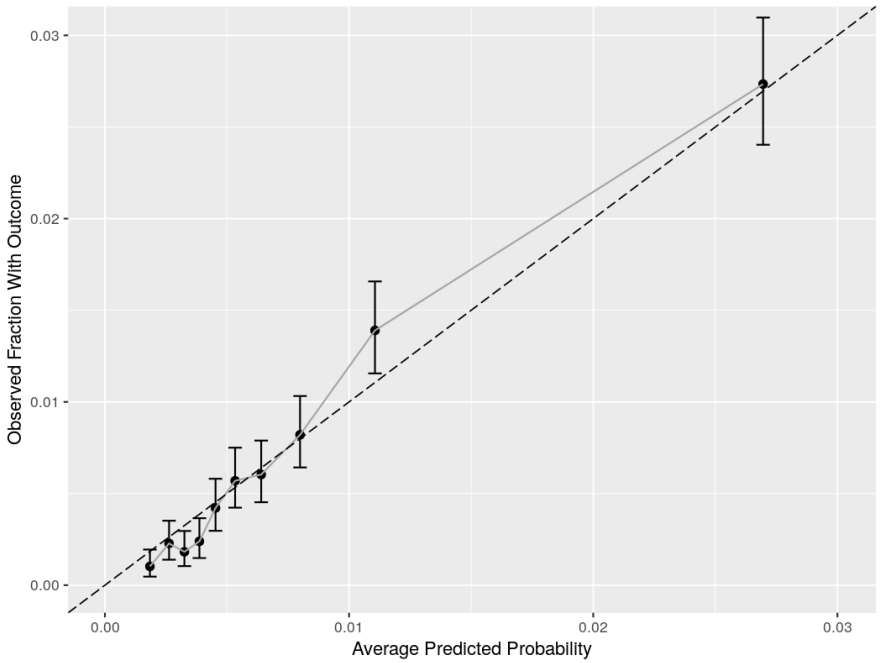


Precision recall plot

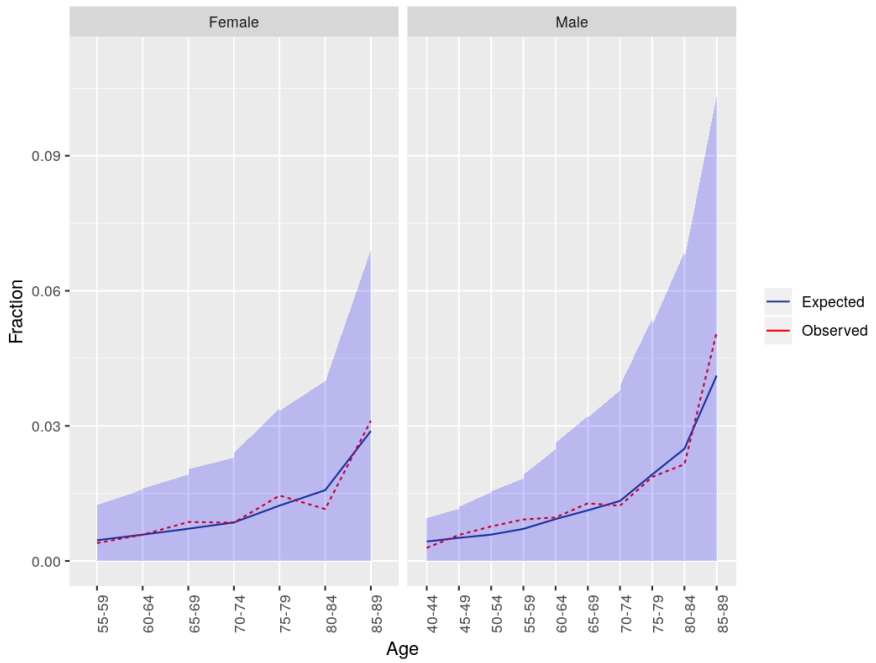


View Interactively via Shiny

Calibration Plot



Demographic Plot



Demo: Using Atlas to design a PLP study

- How to add Target and Outcome cohorts
- How to define population settings
- How to define covariate settings
- How to define model settings
- How to save/load/copy/delete PLP analysis
- How to download R package for running study

Demo: Building the R package

- How to open the R package in R studio
- Details about files to edit
- How to build the package
- How to run the package

Demo: Viewing the Shiny App

- How to view results interactively
- How to view settings
- How to view performance
- How to view model
- How to view log

Demo: Creating the validation package and adding to github

- How to create the validation package
- How to add a package to github for external validation

Demo: Creating the journal paper template

- How to convert the results into a template journal paper document

Questions?

Today's Agenda

Time	Topic
8:45 – 9:00	Get settled, get laptops ready
9:00 – 10:00	Exercise: Selection of prediction problem
10:00 – 10:45	Presentation: What is Patient-Level Prediction
10:45 – 11:00	Break
11:00 – 11:45	Presentation: Learning the OHDSI Patient-Level Prediction Framework
11:45 – 12:30	Presentation: Overview of the TRIPOD Statement Exercise: Applying TRIPOD to CHADS2
12:30 – 13:15	Lunch
13:15 – 15:15	Guided tour through implementing patient-level prediction
15:15 – 15:30	Break
15:30 – 16:45	Exercise: Design and implement your own patient-level prediction
16:45 – 17:00	Lessons Learned and Feedback

Transparent **R**eporting of a multivariable prediction model for **I**ndividual **P**rognosis **O**r **D**iagnosis (TRIPOD)

Ross Williams

Erasmus MC

OHDSI

Observational Health Data Sciences and Informatics

Agenda

1. Basics of good reporting for prediction models

2. Review of the TRIPOD Statement

3. Small group discussion of sample paper

4. Large group summary of small group findings

Requirements for clinical implementation

- Clinical setting
  - Clinician should be able to identify for who, predicting what, in what time-at-risk
- Evidence of performance
  - Well calibrated?
  - Good discrimination?

Requirements for clinical implementation

Most models reported in the literature do not provide enough information to impact clinical practice

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration

Karel G.M. Moons, PhD; Douglas G. Altman, DSc; Johannes B. Reitsma, MD, PhD; John P.A. Ioannidis, MD, DSc; Petra Macaskill, PhD; Ewout W. Steyerberg, PhD; Andrew J. Vickers, PhD; David F. Ransohoff, MD; and Gary S. Collins, PhD

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) Statement includes a 22-item checklist, which aims to improve the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. The TRIPOD Statement aims to improve the transparency of the reporting of a prediction model study regardless of the study methods used. This explanation and elaboration document describes the rationale; clarifies the meaning of each item; and discusses why transparent reporting is important, with a view to assessing risk of bias and clinical usefulness of the prediction model. Each checklist item of the TRIPOD Statement is explained in detail and accompanied by published examples of good reporting. The document also provides a valuable reference of issues to consider when designing, conducting, and analyzing prediction model studies. To aid the editorial process and help peer reviewers and, ultimately, readers and systematic reviewers of prediction model studies, it is recommended that authors include a completed checklist in their submission. The TRIPOD checklist can also be downloaded from www.tripod-statement.org.

Ann Intern Med. 2015;162:W1-W73. doi:10.7326/M14-0698 **www.annals.org**
For author affiliations, see end of text.
For members of the TRIPOD Group, see the **Appendix**.

Developed by 25+ member committee of prediction modeling experts

Reduced from 76 to 22 items

1 Title

Concise summary of the model

"**Development and validation** of a clinical score to **estimate the probability of coronary artery disease** in **men and women presenting with suspected coronary disease**"

3 Background and Objectives

What was the goal for developing this model?

"The aim of this study was to **develop and validate** a **clinical prediction rule** in **women presenting with breast symptoms**, so that a **more evidence based approach to referral**—which would include urgent referral under the 2 week rule—could be implemented as part of clinical practice guidance."

4 Methods - Source of Data

Gives an indication of applicability and quality of the data

"The population based sample used for this report included **2489 men and 2856 women 30 to 74 years old** at the time of their **Framingham Heart Study** examination in 1971 to 1974. Participants attended either the 11th examination of the original Framingham cohort or the initial examination of the Framingham Offspring Study. Similar research protocols were used in each study, and **persons with overt coronary heart disease at the baseline examination were excluded**."

6 Methods- Outcome

What was predicted and how was it measured?

"Breast Cancer Ascertainment: **Incident diagnoses of breast cancer** were **ascertained by self-report on biennial follow up questionnaires** from 1997 to 2005. We learned of deaths from family members, the US Postal Service, and the National Death Index. We identified 1084 incident breast cancers, and **1007 (93%) were confirmed by medical record or by cancer registry data** from 24 states in which 96% of participants resided at baseline."

7 Methods- Predictors

What was used to inform the model? When was the data collected?

"The following data were extracted for each patient: **gender, aspartate aminotransferase in IU/L, alanine aminotransferase in IU/L, aspartate** aminotransferase/ alanine aminotransferase ratio, total bilirubin (mg/dl), albumin (g/dl), transferrin saturation (%), mean corpuscular volume (µm3), platelet count (× 103/mm3), and prothrombin time(s). . . . **All laboratory tests were performed within 90 days before liver biopsy**. In the case of repeated test, the results closest to the time of the biopsy were used. No data obtained after the biopsy were used.

10 Methods - Statistics

What type of model was used and how was performance assessed?

"We used the **Cox proportional hazards model** in the derivation dataset to estimate the coefficients associated with each potential risk factor [predictor] for the first ever recorded diagnosis of cardiovascular disease for men and women separately."

"We assessed the predictive performance of the QRISK2- 2011 risk score on the THIN cohort by **examining measures of calibration and discrimination**… **Calibration** of the risk score predictions was assessed by **plotting observed proportions versus predicted probabilities** and by calculating the calibration slope… **Discrimination** … quantified by **calculating the area under the receiver operating characteristic curve statistic**; a value of 0.5 represents chance and 1 represents perfect discrimination."

15 Results – Model Specification

What were the predictors and how were they used to inform the final prediction?

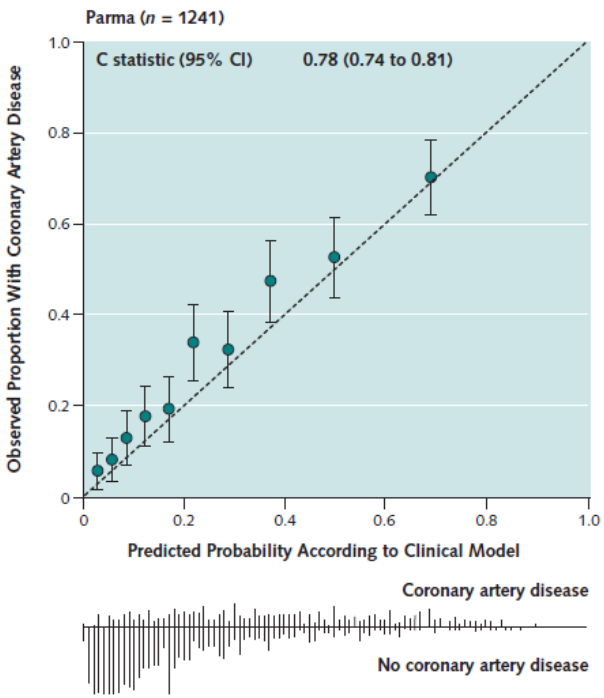
Table 12. Example Table: Presenting the Full Prognostic (Survival) Model, Including the Baseline Survival, for a Specific Time Point*

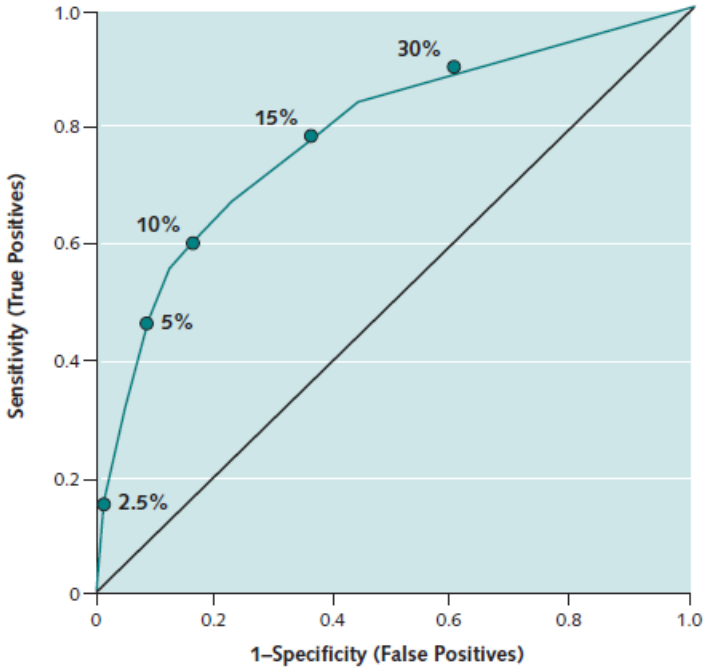
	β Coefficient	SE	*P* Value
Age	0.15052	0.05767	0.009
Age^2	−0.00038	0.00041	0.35
Male sex	1.99406	0.39326	0.0001
Body mass index	0.01930	0.01111	0.08
Systolic blood pressure	0.00615	0.00225	0.006
Treatment for hypertension	0.42410	0.10104	0.0001
PR interval	0.00707	0.00170	0.0001
Significant cardiac murmur	3.79586	1.33532	0.005
Heart failure	9.42833	2.26981	0.0001
Male sex × age^2	−0.00028	0.00008	0.0004
Age × significant murmur	−0.04238	0.01904	0.03
Age × prevalent heart failure	−0.12307	0.03345	0.0002

From reference 402.
* $S_0(10) = 0.96337$ (10-year baseline survival). β values are expressed per 1-unit increase for continuous variables and for the condition present in dichotomous variables.

16 Results - Performance

How well did the model perform based on the specified metrics?





Small Group Discussion

- Review "Validation of Clinical Classification Schemes for Predicting Stroke Results From the National Registry of Atrial Fibrillation" Gage et al.
- Group assignment for filling in the TRIPOD table
- Grade each item:
  - A: completely fulfills the requirement
  - C: partially fulfills the requirement
  - F: does not fulfill the requirement
- Take about 20 minutes

Quiz time!

- All questions framed as whether the paper we read meets one specific part of the Tripod statement
- The quiz consists of:
  - 10 multiple choice questions
  - The faster you answer correctly, the better your score
- There is a prize…

Online training environment

We will be working in R Studio and Atlas:

RStudio:
https://rstudio.plp.ohdsitutorials.amazingawsdemos.com/

Atlas:
https://plp.ohdsitutorials.amazingawsdemos.com/

Username: **userX** (X you can find on agenda)
Password: **Password1**
Please use Chrome for the exercises.

Lunch Time

Today's Agenda

Time	Topic
8:45 – 9:00	Get settled, get laptops ready
9:00 – 10:00	Exercise: Selection of prediction problem
10:00 – 10:45	Presentation: What is Patient-Level Prediction
10:45 – 11:00	Break
11:00 – 11:45	Presentation: Learning the OHDSI Patient-Level Prediction Framework
11:45 – 12:30	Presentation: Overview of the TRIPOD Statement Exercise: Applying TRIPOD to CHADS2
12:30 – 13:15	Lunch
13:15 – 15:15	Guided tour through implementing patient-level prediction
15:15 – 15:30	Break
15:30 – 16:45	Exercise: Design and implement your own patient-level prediction
16:45 – 17:00	Lessons Learned and Feedback

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Exercise:

Guided tour through implementing patient-level prediction

Task (Modified CHADS2 model)

In target population (**PLP training: T : patients newly diagnosed with Atrial fibrillation)** predict who will develop outcome (**PLP training: O - hospitalized ischemic stroke events)** during the period from 0 days from cohort start date to 1000 days.

Example

We implemented three models in OPTUM for the prediction problem:

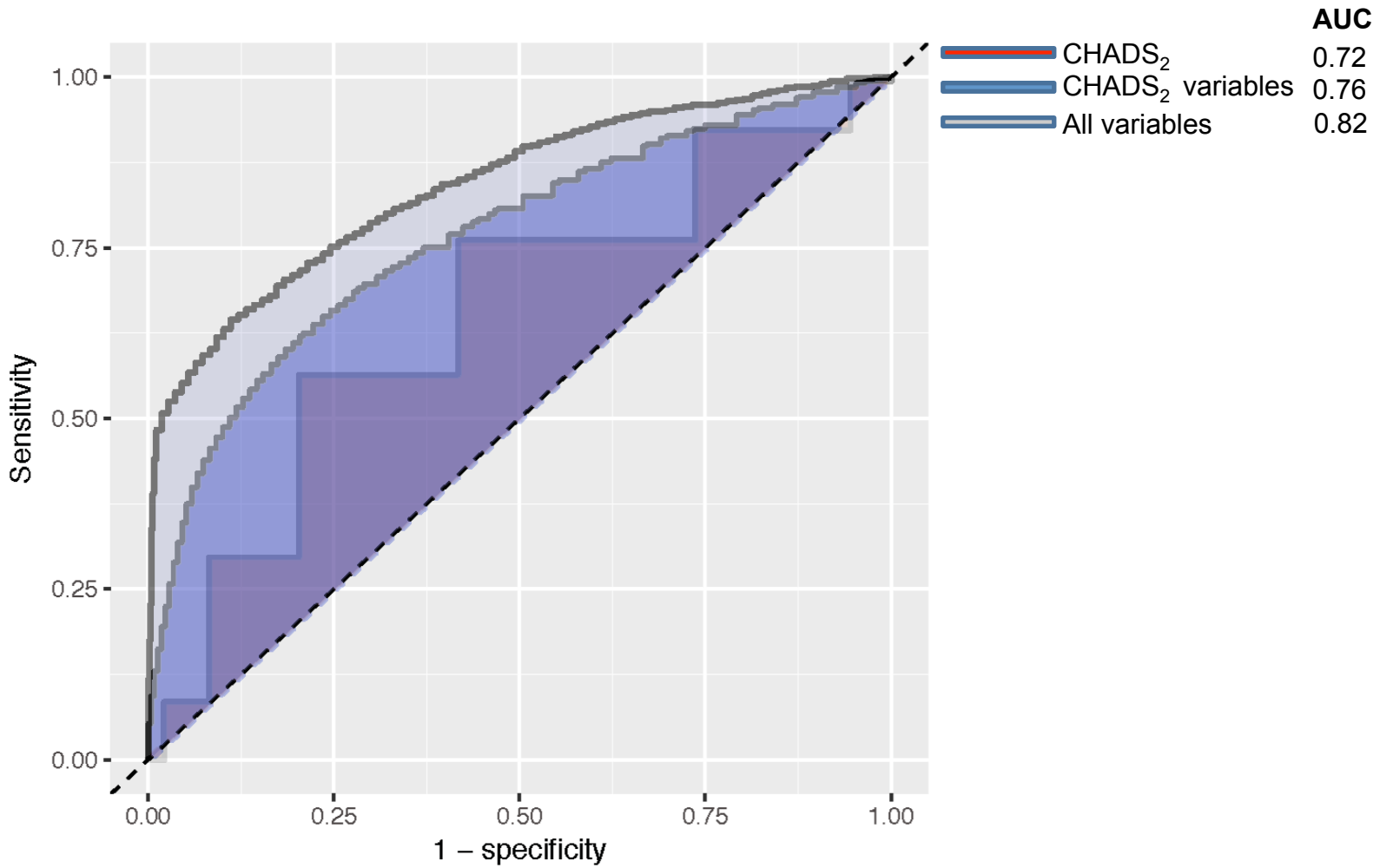
1. CHAD2 model
2. PLP model using 5 CHAD2 variables (and descendants)
3. PLP model using all variables

DEMO

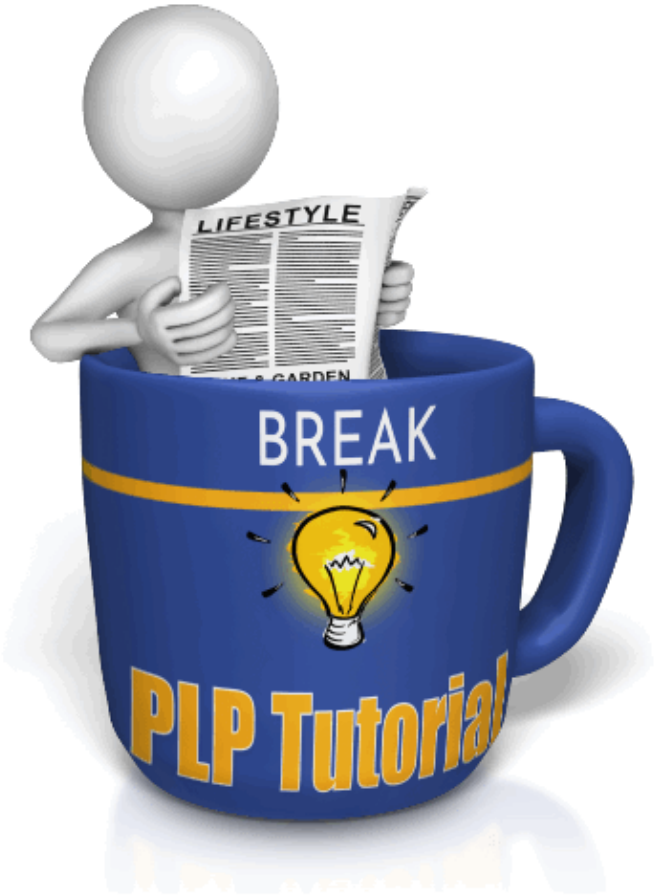
OHDSI

Observational Health Data Sciences and Informatics

Predicting Stroke in Patients with Atrial Fibrillation: OPTUM results



Let’s take a 15 min break



Today's Agenda

Time	Topic
8:45 – 9:00	Get settled, get laptops ready
9:00 – 10:00	Exercise: Selection of prediction problem
10:00 – 10:45	Presentation: What is Patient-Level Prediction
10:45 – 11:00	Break
11:00 – 11:45	Presentation: Learning the OHDSI Patient-Level Prediction Framework
11:45 – 12:30	Presentation: Overview of the TRIPOD Statement Exercise: Applying TRIPOD to CHADS2
12:30 – 13:15	Lunch
13:15 – 15:15	Guided tour through implementing patient-level prediction
15:15 – 15:30	Break
15:30 – 16:45	Exercise: Design and implement your own patient-level prediction
16:45 – 17:00	Lessons Learned and Feedback

OHDSI

Observational Health Data Sciences and Informatics

Exercise:

Design and implement your own patient-level prediction

Exercise

- Read the Patient-Level Prediction Vignette
- You can define new cohorts in Atlas or use those that are there

Option 1: Make a study package through Atlas

Option 2: Make your own script by following the vignette

Things to Explore

1) What is the effect of the length of the time-at-risk period on performance?
2) What is the difference in performance of the algorithms?

Hints:

- sample your cohorts to max 10.000 patients to improve speed.
- start with regularized regression.

Today's Agenda

Time	Topic
8:45 – 9:00	Get settled, get laptops ready
9:00 – 10:00	Exercise: Selection of prediction problem
10:00 – 10:45	Presentation: What is Patient-Level Prediction
10:45 – 11:00	Break
11:00 – 11:45	Presentation: Learning the OHDSI Patient-Level Prediction Framework
11:45 – 12:30	Presentation: Overview of the TRIPOD Statement Exercise: Applying TRIPOD to CHADS2
12:30 – 13:15	Lunch
13:15 – 15:15	Guided tour through implementing patient-level prediction
15:15 – 15:30	Break
15:30 – 16:45	Exercise: Design and implement your own patient-level prediction
16:45 – 17:00	Lessons Learned and Feedback

Lessons learned and feedback

Lessons Learned

Learned the PLP Dance

Educated Fortune Teller

What's Next?

When you write your JAMA publication;

1. Follow the TRIPOD Statement.
2. Cite our work:

JAMIA
A SCHOLARLY JOURNAL OF INFORMATICS IN HEALTH AND BIOMEDICINE

Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data

Jenna M Reps, Martijn J Schuemie, Marc A Suchard, Patrick B Ryan, Peter R Rijnbeek

Journal of the American Medical Informatics Association, Volume 25, Issue 8, August 2018, Pages 969–975, https://doi.org/10.1093/jamia/ocy032
Published: 27 April 2018 **Article history** ▾

PDF Split View Cite Permissions Share ▾

Abstract

Objective

To develop a conceptual prediction model framework containing standardized steps and describe the corresponding open-source software developed to consistently implement the framework across computational environments and observational healthcare databases to enable model sharing and reproducibility.

R-package

www.github.com/OHDSI/PatientLevelPrediction

- Vignettes
- Videos
- Online training material

Book-of-OHDSI
https://ohdsi.github.io/TheBookOfOhdsi/

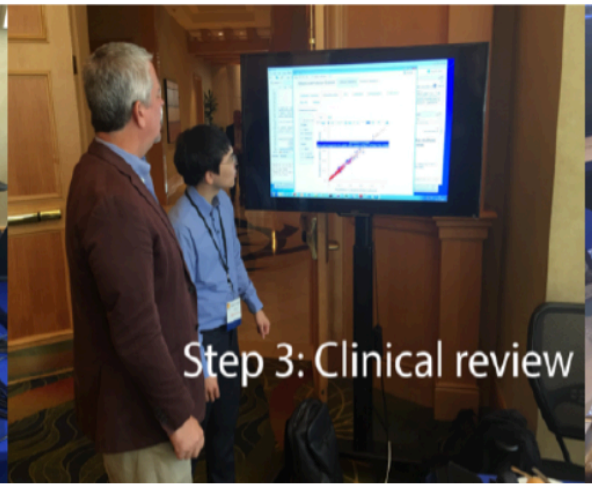
Study Results
www.data.ohdsi.org

Large-Scale Patient-Level Prediction not the Future!

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Patient-Level Prediction Team Work



Join the PLP Community

- Monthly meetings of PLP WG
- Researchers Forum (tag patientprediction)
- Become an active developer: add your own algorithms and other features

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
- Transportability
- Temporal information
- Textual information
- Deep learning
- Ensemble training
- Learning Curves

Clinical impact for the patient

- How to assess?

Tool Development

- Model Library
- Results viewer improvements.

Thank you!

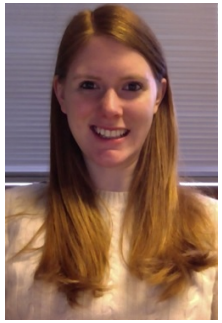
OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

This tutorial would not have been possible without the contribution of many collaborators in the OHDSI Community

aws

We like to thank Amazon Web Services for their valuable technical support and resources

Faculty

Peter Rijnbeek Erasmus MC	Ross Williams Erasmus MC	Jenna Reps Janssen R&D	Patrick Ryan Janssen R&D
			

https://www.ohdsi.org/who-we-are/collaborators/

Tutorial improvement

We like to hear your feedback on this course:

- What went well?
- What did not?
- What do you like to see added?
- You can give your feedback on the evaluation form:

https://bit.ly/2EeSIpC

Questions? Drop us an email

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