OHDSI Tutorial:

Patient-Level Prediction

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.

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

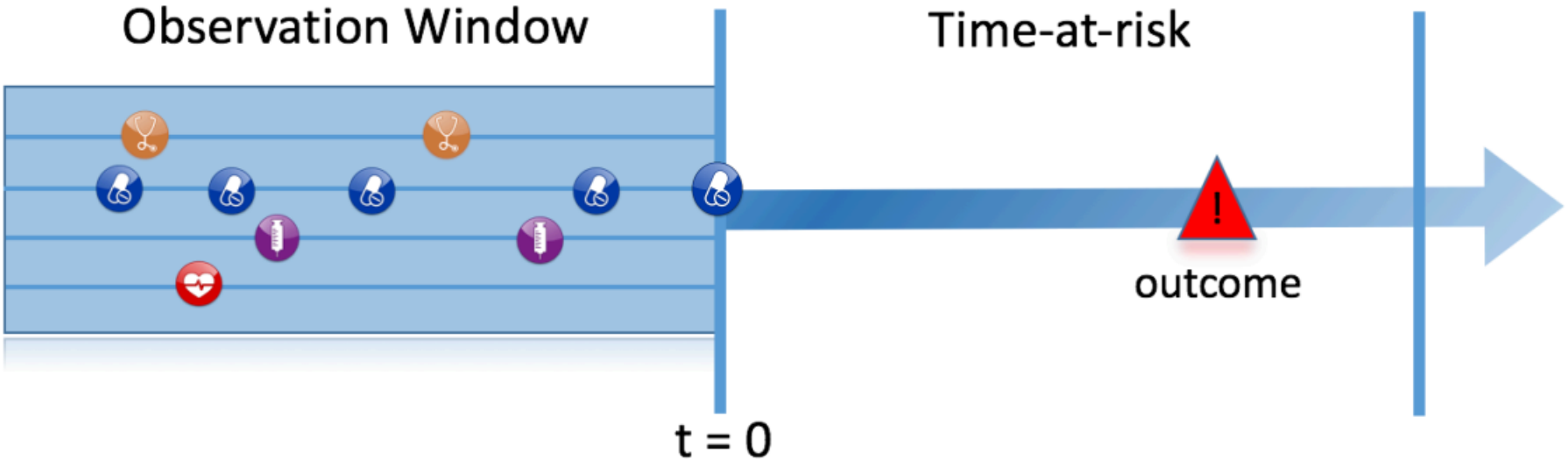
OHDSI aims to develop a systematic process to learn and evaluate large-scale patient-level prediction models using observational health data in a data network

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.

Problem definition

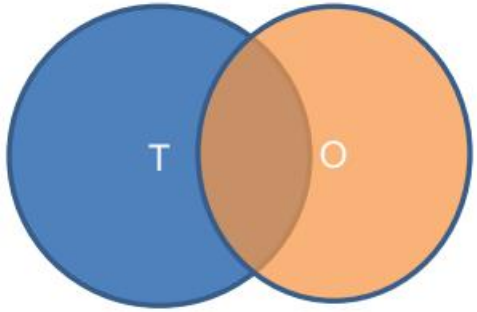


Among a target population (T), we aim to predict which patients at a defined moment in time (t=0) will experience some outcome (O) during a time-at-risk Prediction is done using only information about the patients in an observation window prior to that moment in time.

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?	

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.

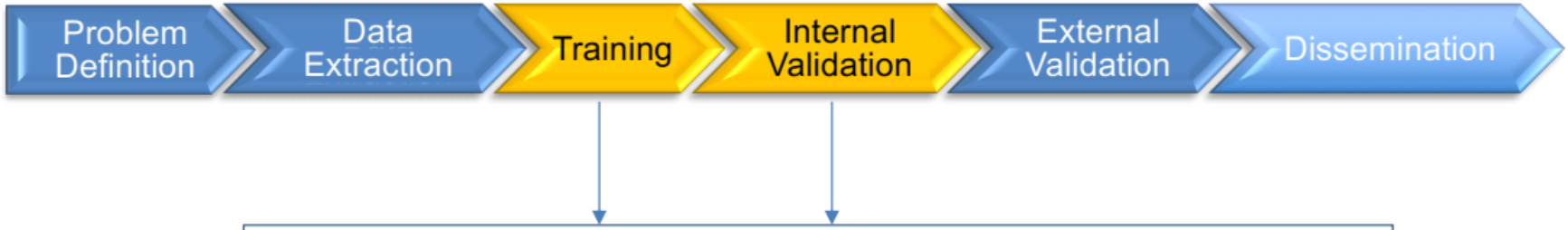
Prediction Model Development

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.

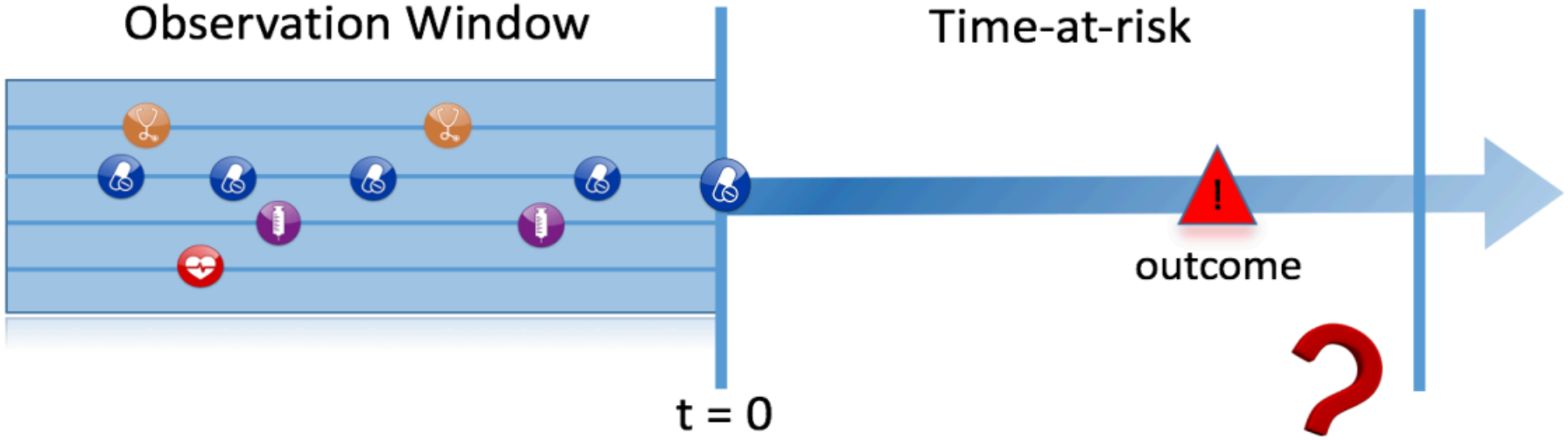
Prediction Model Development



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)

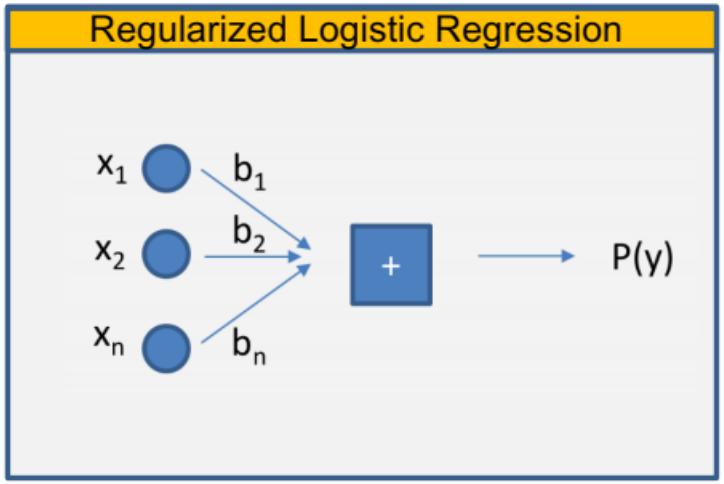
Model Training

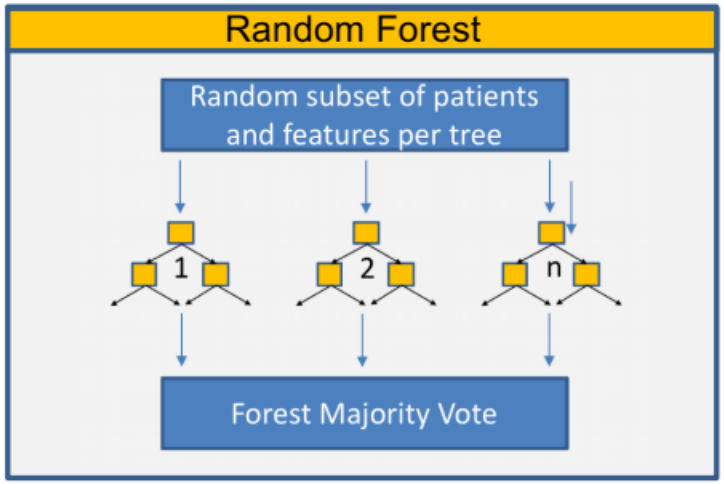


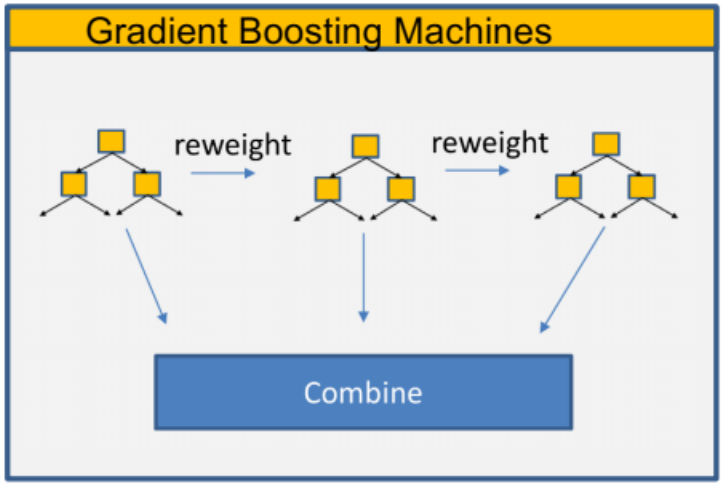
1. Which models?

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.

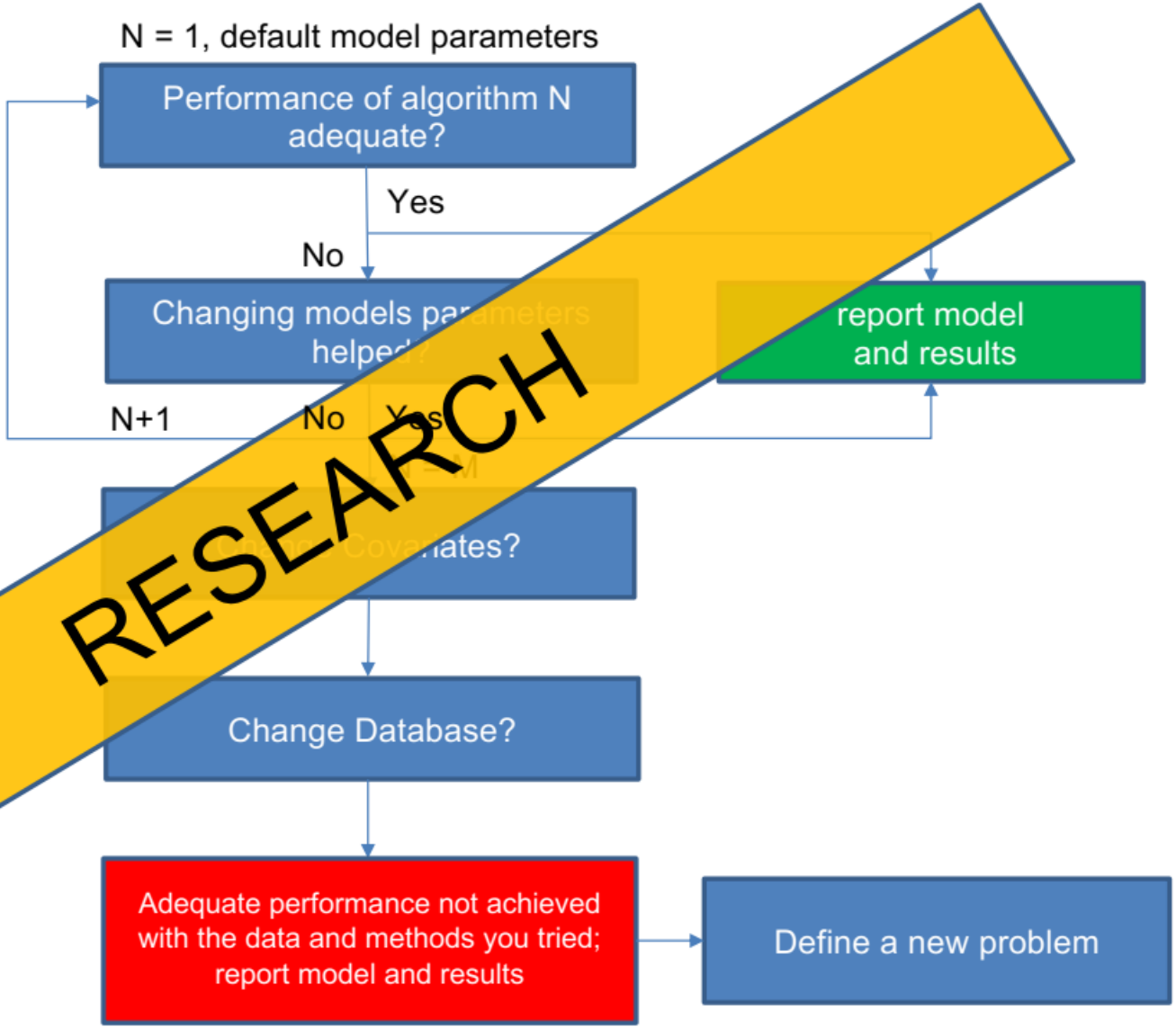
It is common in machine learning to try multiple models and find one that works best for that particular problem.

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.	…



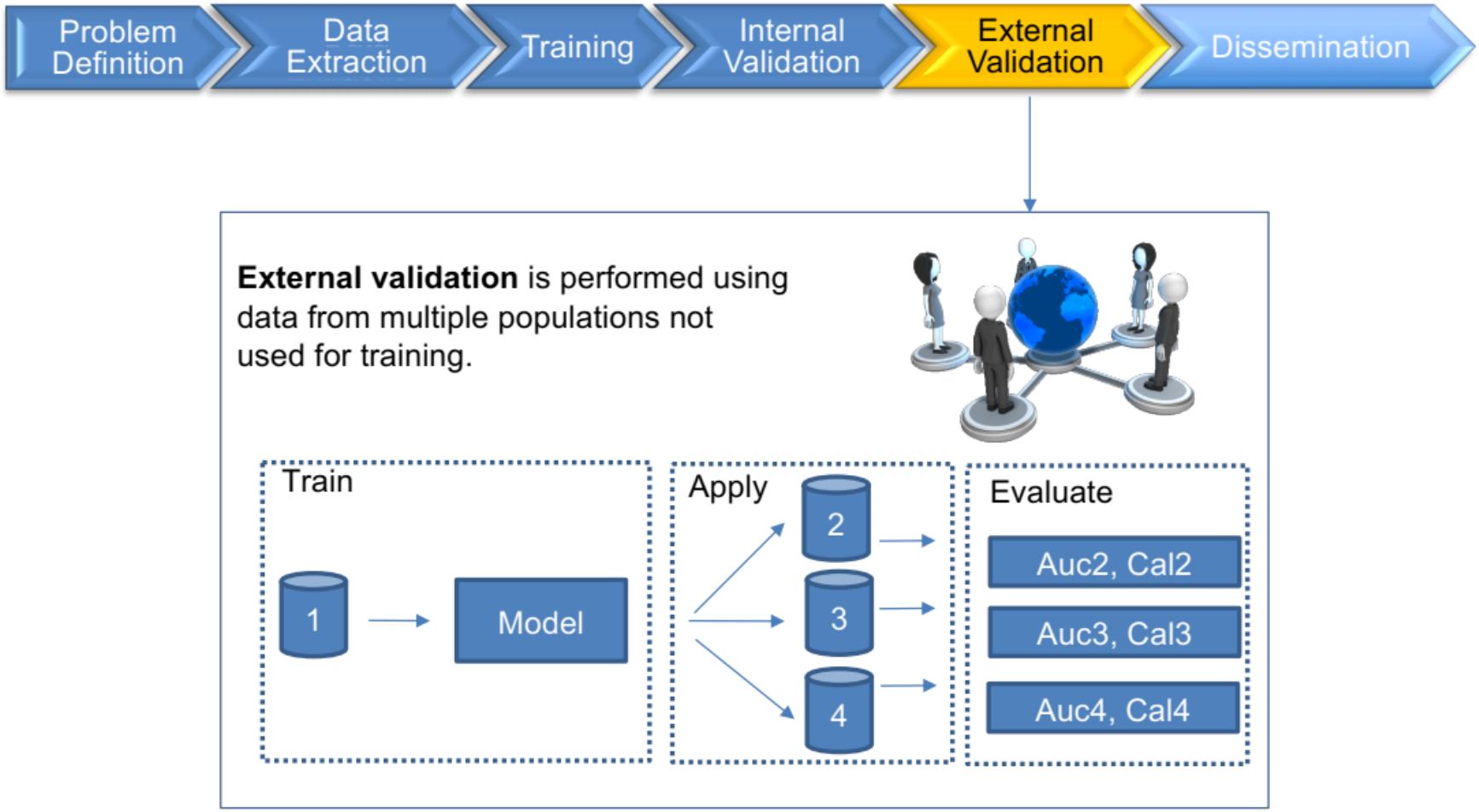
Model Validation

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

External Validation



Dissemination

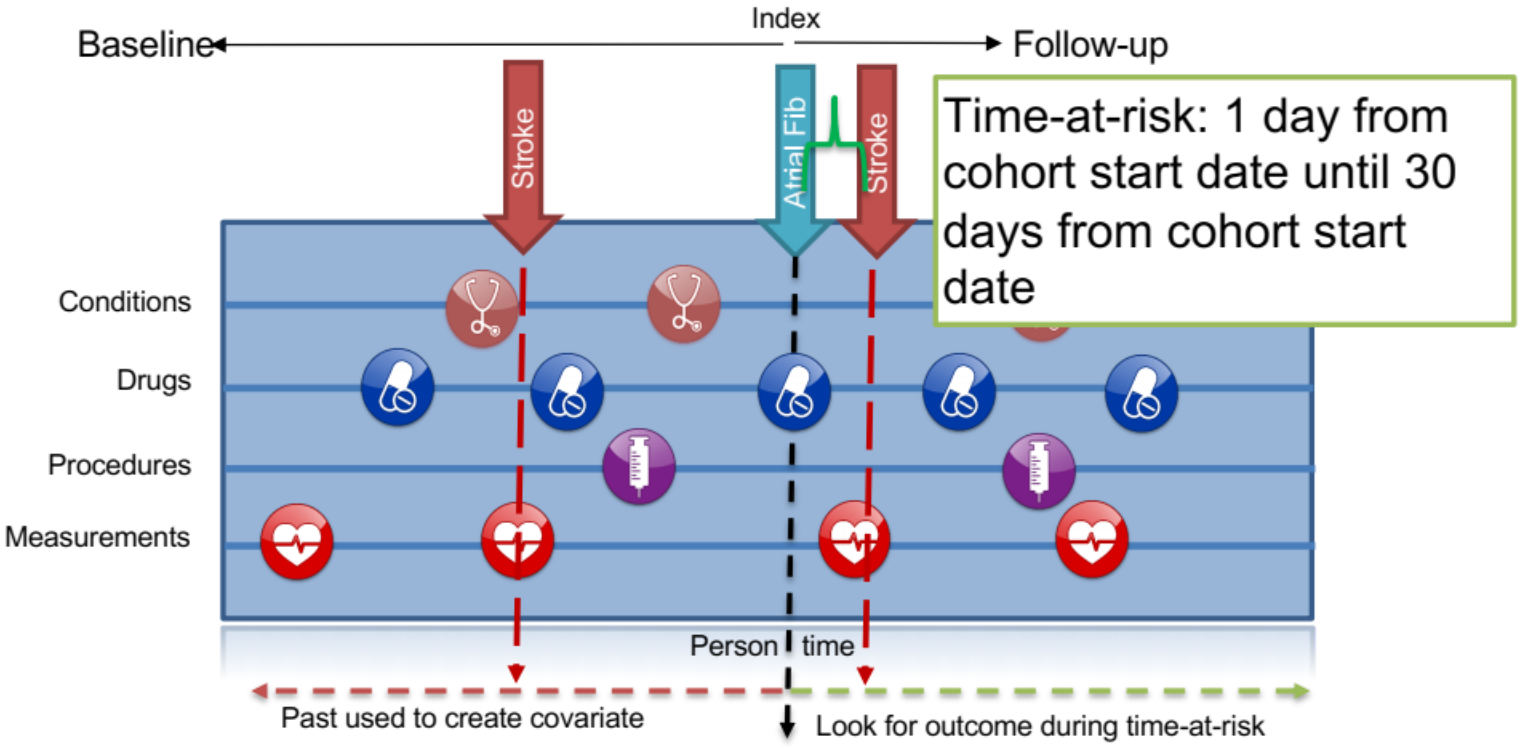
Dissemination of study results should follow the minimum requirements as stated in the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement [1].

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

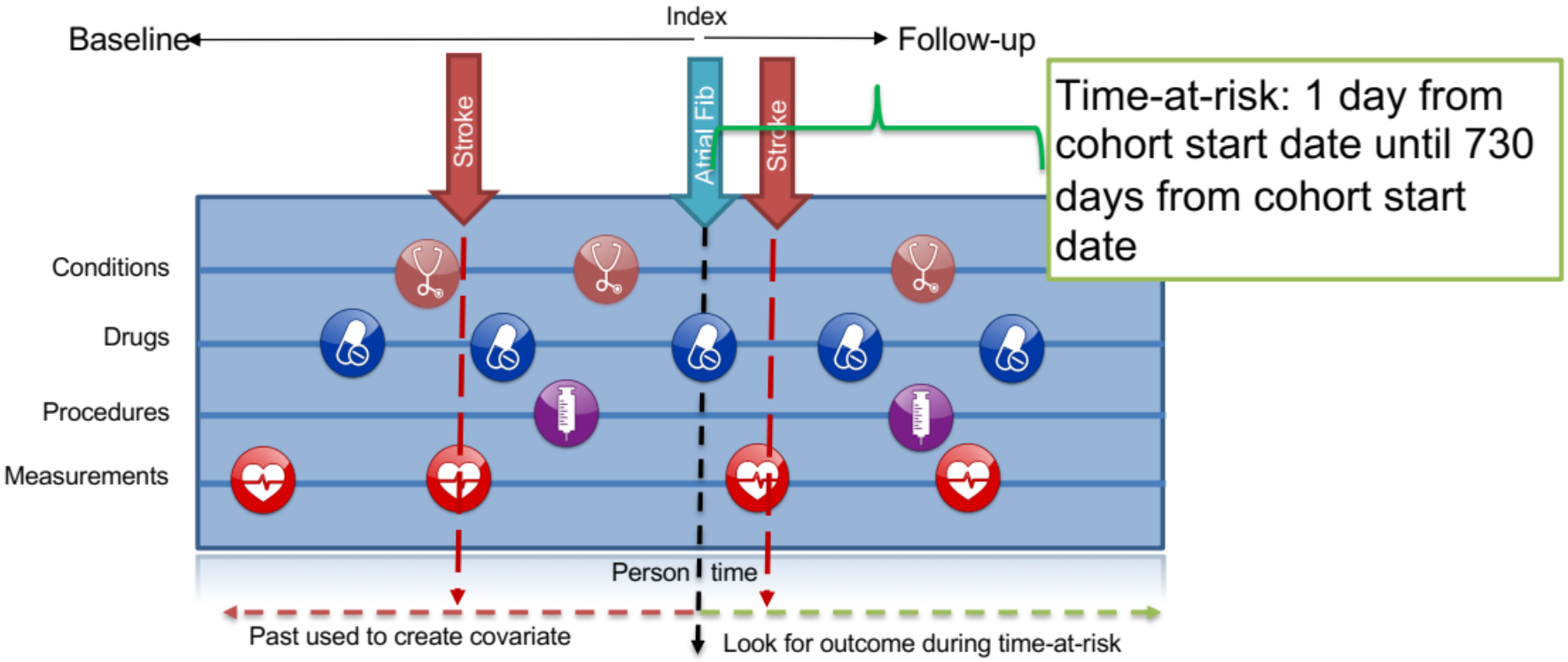
Website to share protocol, code, models and results for all databases

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

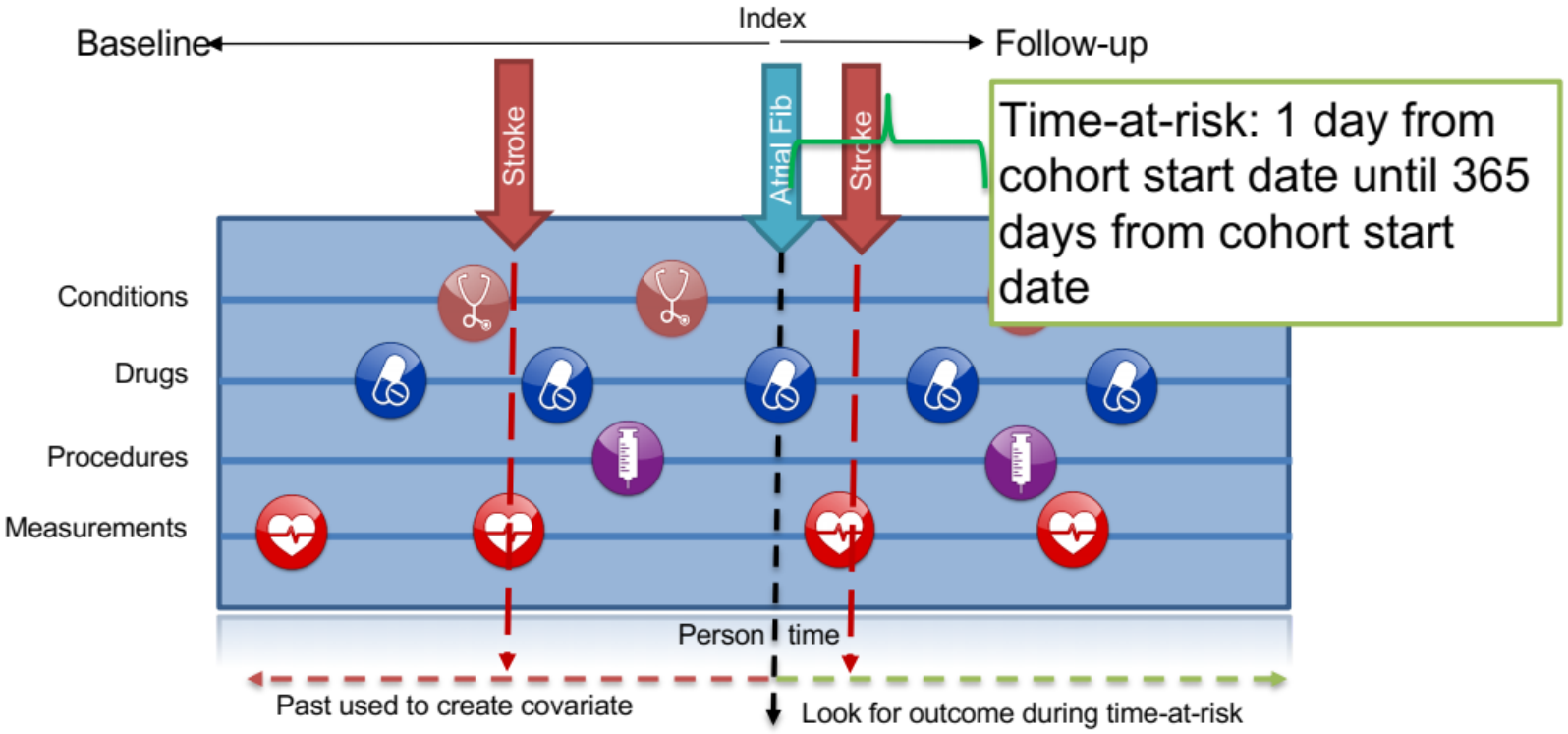
Extracting Labelled Data



Extracting Labelled Data



Extracting Labelled Data



Extracting Labelled Data

Each person corresponds to a row

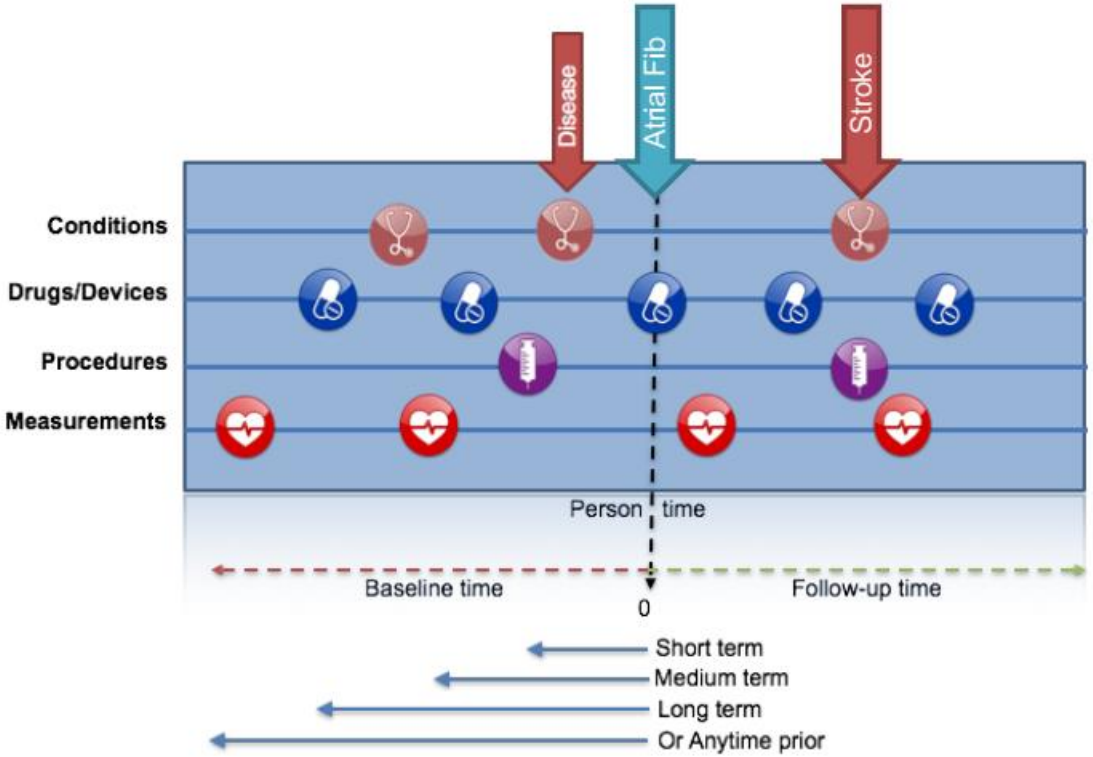
Labelled classification data

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., the atrial fibrillation)



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*	*2013-10-04*	*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? **[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
105454	1	2012-08-12	0	2009
4346356	1	2011-07-05	1	4056
…	…	…		

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~~	~~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? **[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~~	~~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? **[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

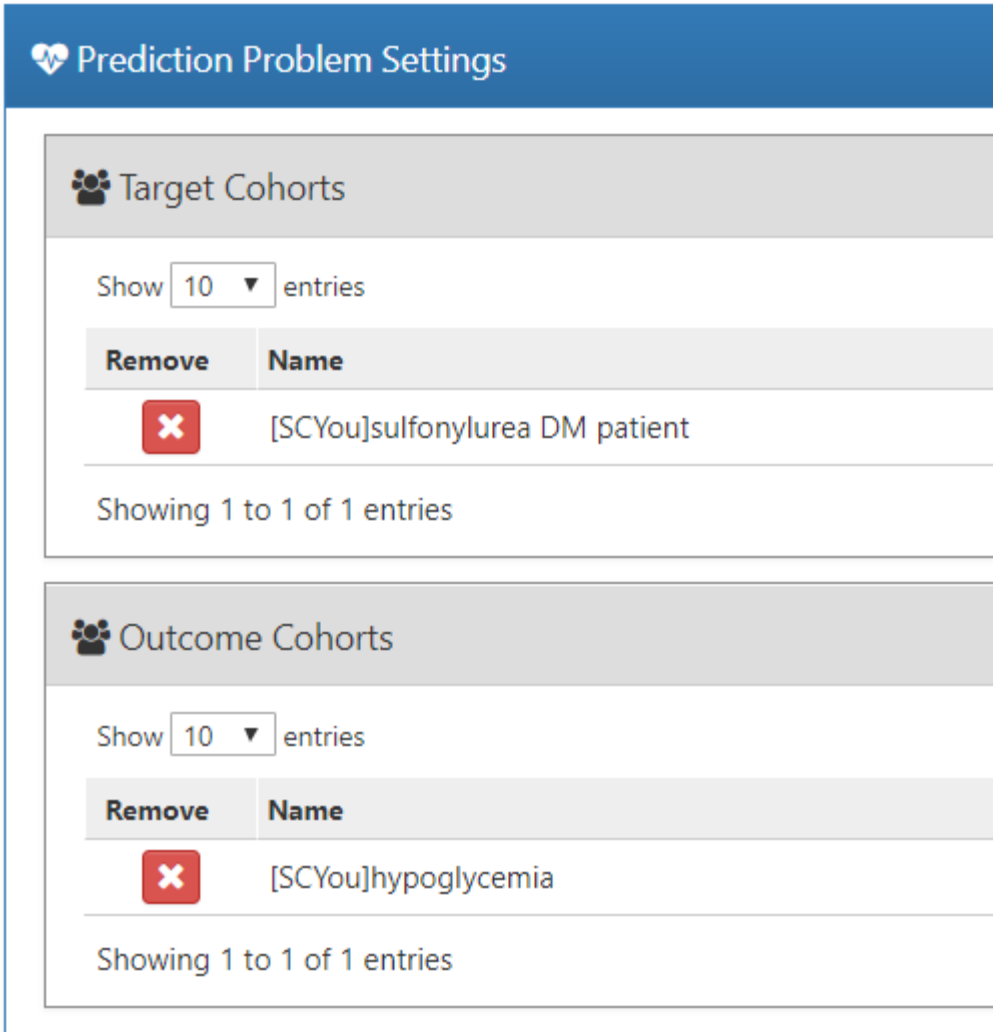
Subject_id	Cohort_id	Cohort_start_date	Outcome during TAR	Follow-up observation
3454102	1	2012-01-02	0	50
105454	1	2012-08-12	0	1082
1554	1	2009-05-05	0	366
56566	1	2011-07-05	0	480
4346356	1	2011-07-05	1	40
342424	1	2010-01-01	1	500
…	…	…		

- 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]**

Design an machine-learning algorithm

Input parameter	Design choice
Target cohort (T)	Sulfonylurea user
Outcome cohort (O)	Hypoglycemia
Time-at-risk	1~90 days
Model specification	

Prediction Problem Settings



Model and Covariates Settings

Model Settings

Show 10 entries

Remove	Model	Options
✖	RandomForestSettings	{"mtries":[-1],"ntrees":[500],"maxDepth":[4,10,17],"varImp":[true],"seed":null}
✖	MLPSettings	{"size":[4],"alpha":[0.00001],"seed":null}
✖	LassoLogisticRegressionSettings	{"variance":0.01,"seed":null}

Showing 1 to 3 of 3 entries

Covariate Settings

Column visibility | Copy | CSV | Show 10 entries

Remove	Options
✖	DemographicsGender, DemographicsAgeGroup, DemographicsRace, ConditionGroupEraLongTerm, DrugGroupEraLongTerm, ProcedureOccurrenceLongTerm (+1 more covariate settings)

Showing 1 to 1 of 1 entries

Population Settings