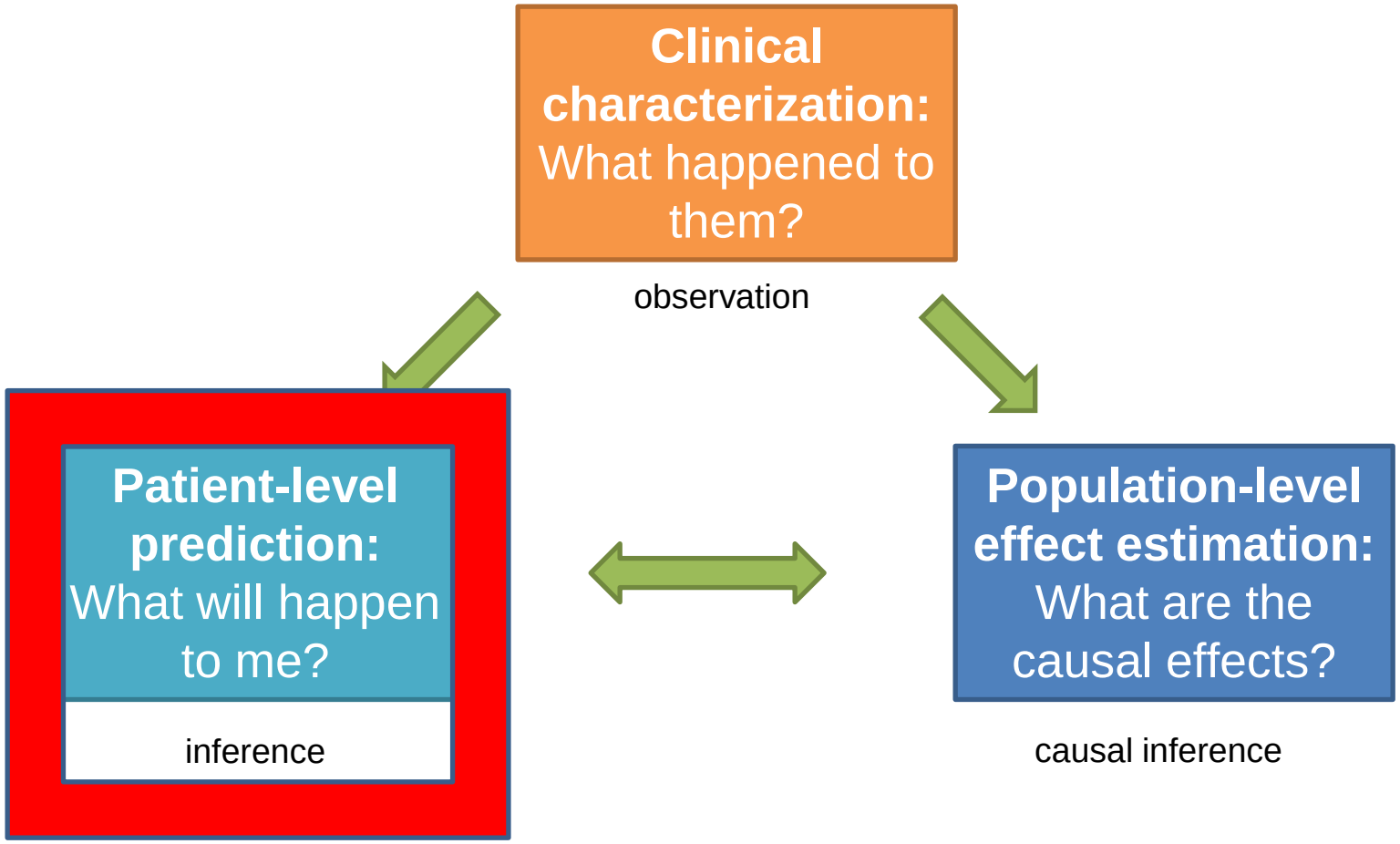
Prediction Tutorial

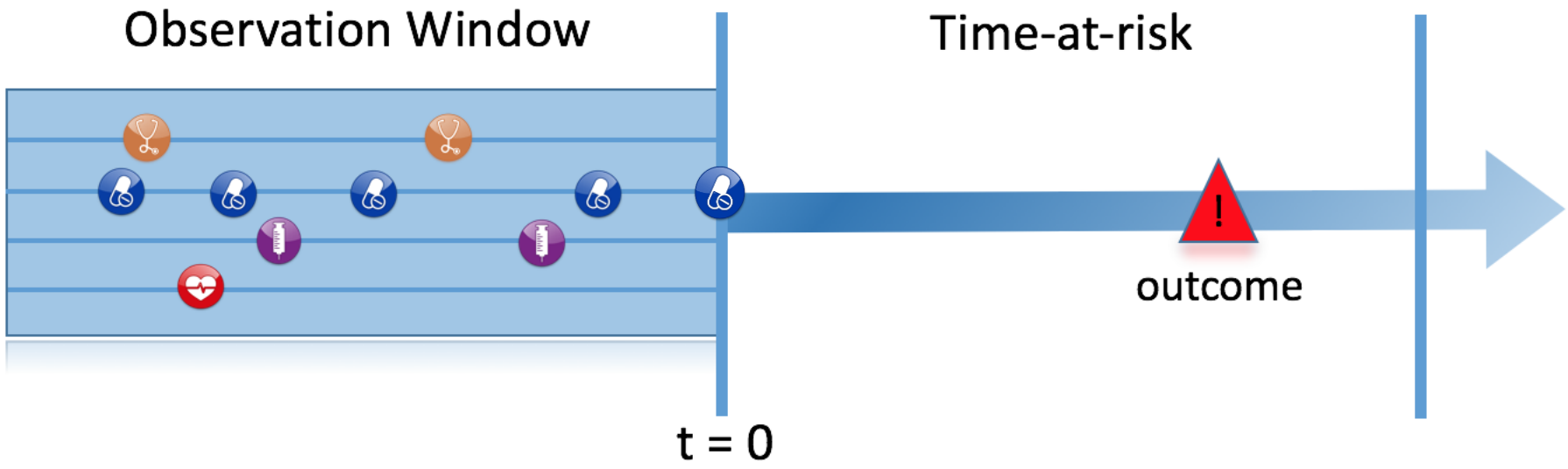
Cynthia Yang, PhD Student, Erasmus MC
Chungsoo Kim, PhD Student, Ajou University

Amongst new users of T: lisinopril, what is risk of O1: angioedema and O2: acute myocardial infarction during 1-year post-exposure?

Complementary evidence to inform the patient journey



Problem definition



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

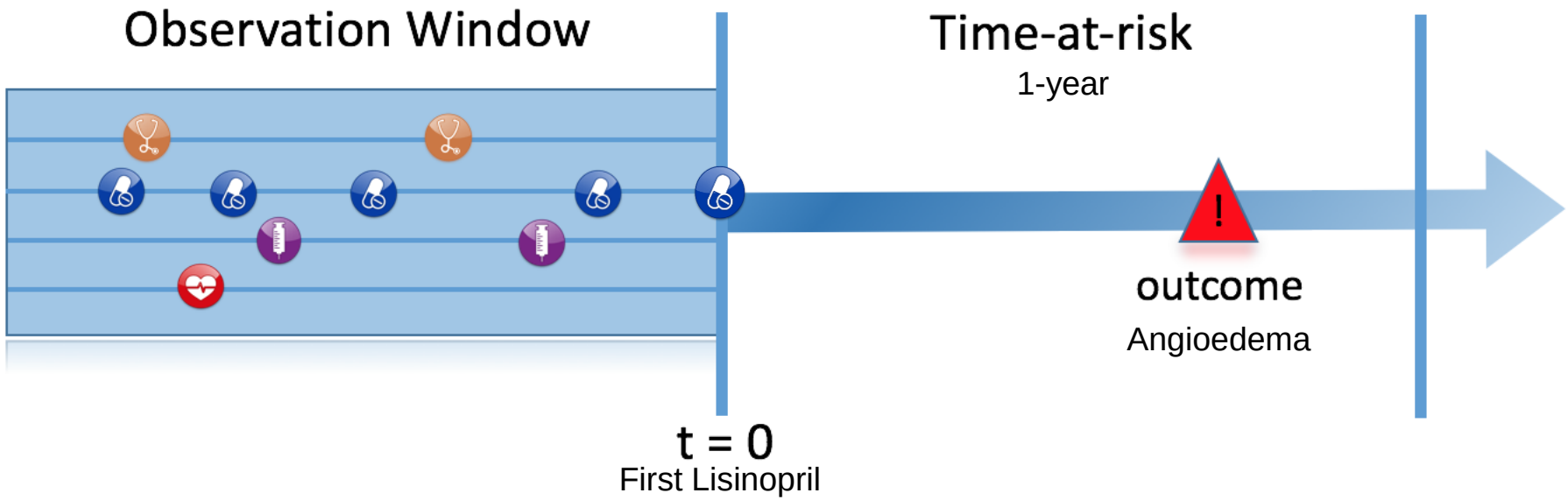
Prediction task specification

Component	Description
Target population (T):	Who do you want to do the prediction for?
Outcome (O):	What are you predicting?
Time-at-risk (TAR):	When are you predicting?

Prediction task specification

Component	Description	
Target population (T):	Who do you want to do the prediction for?	New users of lisinopril
Outcome (O):	What are you predicting?	Angioedema
Time-at-risk (TAR):	When are you predicting?	1 day to 365 days after lisinopril

Specified Problem definition



Among a **target population (T) of new users of lisinopril**, we aim to predict which patients at the **index date of lisinopril dispensing (t=0)** will experience the **outcome angioedema (O)** during **1 day to 365 days after index**. Prediction is done using only information about the patients in an observation window prior to index.

There are many design choices to make

Data Extraction

- What design? Case control vs cohort
- How to address loss to follow up?
- How far back for covariate lookback?
- How to define T and O

Model Development

- Should we preprocess data?
- What classifier to fit?
- Should we look at prediction diagnostics/learning curves

Model Validation

- What data should we use to validate?
- What metrics to report?

Current Best Practices For Model Design

https://ohdsi.github.io/PatientLevelPrediction/articles/BestPractices.html

Data Extraction

- Use a cohort design
- Exclude those with and without the outcome equally
- Use enough time prior to index for covariate construction (>=180 days)
- Check phenotypes for T and O

Model Development

- Do not do under/over sampling
- Try different classifiers
- Run model design diagnostics for the database (sufficient outcome count)

Model Validation

- Use some form of hold out set for internal validation
- Do external validation (across the OHDSI network)
- Make sure to include calibration and discrimination

PLP Framework Enforces Most…

Data Extraction

- **Use a cohort design**
- **Exclude those with and without the outcome equally**
- **Use enough time prior to index for covariate construction (>=180 days)**
- Check phenotypes for T and O

Model Development

- **Do not do under/over sampling**
- **Try different classifiers**
- **Run model design diagnostics for the database (sufficient outcome count)**

Model Validation

- **Use some form of hold out set for internal validation**
- **Do external validation (across the OHDSI network)**
- **Make sure to include calibration and discrimination**

Let's see prediction in action…

Amongst new users of T: lisinopril, what is risk of O1: angioedema during 1-year post-exposure?

Component	Description
Target population (T):	New users of lisinopril
Outcome (O):	Angioedema
Time-at-risk (TAR):	1 day after first exposure of lisinopril to 365 days after
Model:	LASSO logistic regression + GBM + RF
Covariates:	Age/sex, conditions/drugs groups in prior 1 year, procedures/observations/measurements in prior 1 year
Database:	PharMetrics

Exploring Shiny

We have a shiny for exploring prediction models – let's explore it and see what settings were used, what the model looks like and how well the model did…

Component	Description
Target population (T):	New users of lisinopril
Outcome (O):	Angioedema
Time-at-risk (TAR):	1 day after first exposure of lisinopril to 365 days after
Model:	LASSO logistic regression
Covariates:	Age/sex, conditions/drugs groups in prior 1 year, procedures/observations/measurements in prior 1 year
Database:	PharMetrics

Group activity

- Shiny app: http://52.69.234.188:8080/login

user=ohdsi, password=ohdsi

- [WARNING] It can be unstable with everyone

Group Activity Questions

Component	Description
Target population (T):	New users of lisinopril
Outcome (O):	Myocardial infarction
Time-at-risk (TAR):	1 day after first exposure of lisinopril to 365 days after

1. How many outcomes were used to train the model?
2. What lookback was used for the covariates?
3. Did the model design pass diagnostics?
4. What type of model was developed?
5. What was the top predictor?
6. What was the AUROC for the train, test and CV? Did the model overfit?
7. Was the model well calibrated (why)?
8. What was the attrition?