Activity 11:
Data Preprocessing

ML for Health, Week 11

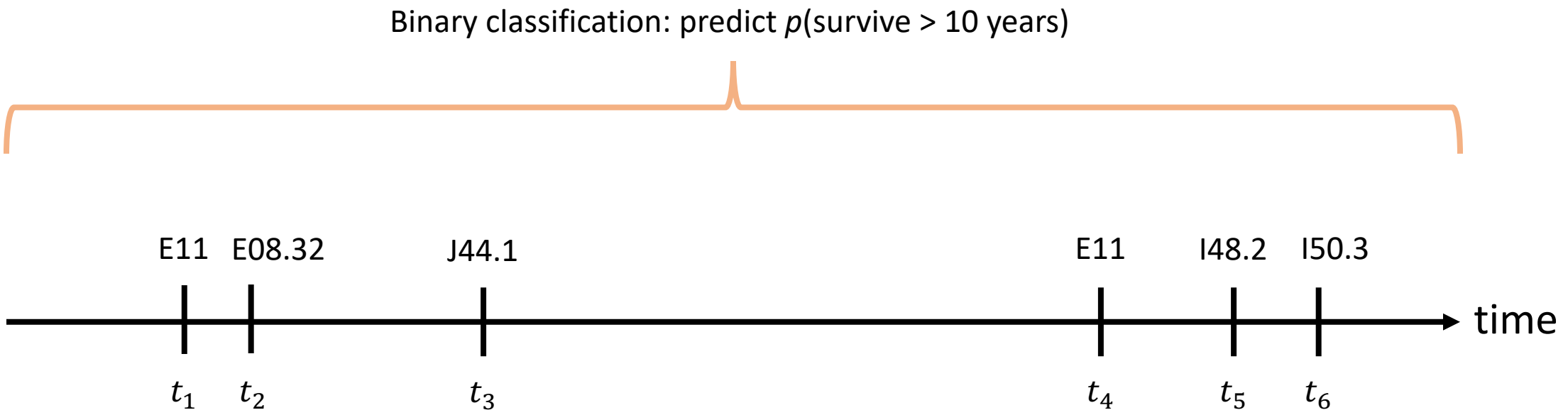
Instructions

- In each of the following scenarios, you'll consider how a specific variety of healthcare data (x) can be used to predict an associated outcome (y)
- Your goal is to (concisely) describe a process by which (x) can be transformed into a fixed length vector suitable for use in a logistic regression or other predictive model
- Simple is good
- When we reconvene, each group will present one of the scenarios

Scenario 1: Diagnosis Codes

Predict patients' 10-year survival probability (y) based on diagnosis codes documented in their chart over the past year (x).

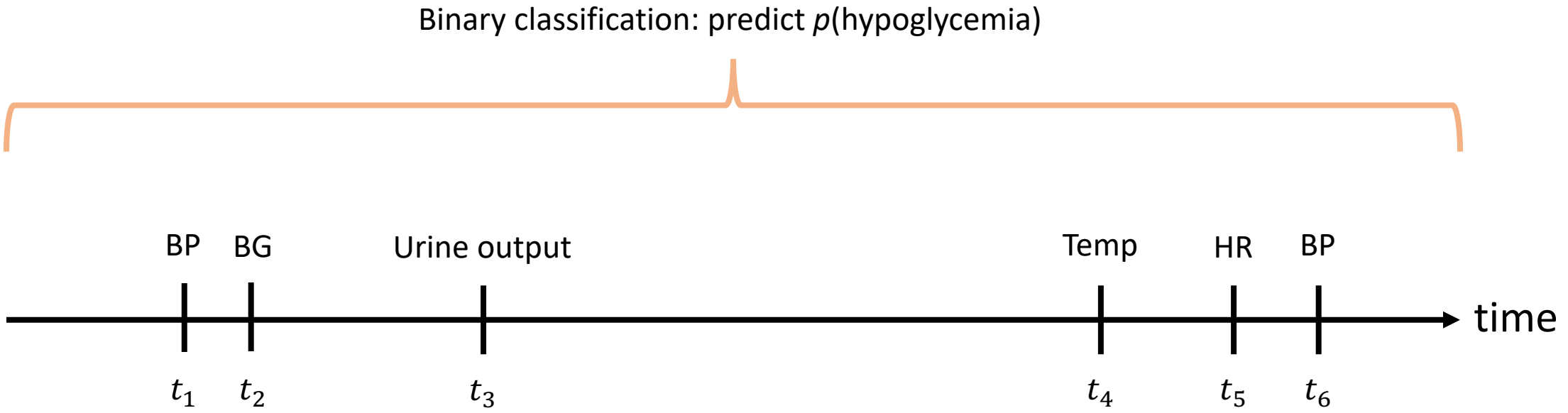
-> How can you transform each sequence into a vector that has the same length for every patient?



Scenario 2: Irregular Measurements

Predict patients' probability of becoming hypoglycemic within the next 6 hours (y) based on vitals and other measurements collected over the last 12 hours (x). Assume that in the average patient, there are many more measurements than shown below.

-> How can you transform each sequence into a vector with no missing values that has the same length for every patient?



Scenario 3: Images with Side Information

Predict whether patients should be referred for diabetic retinopathy (y) based on fundoscopic images *and* demographic information (e.g. age, sex) (x).

-> How can you combine image features with demographic features in a predictive model?