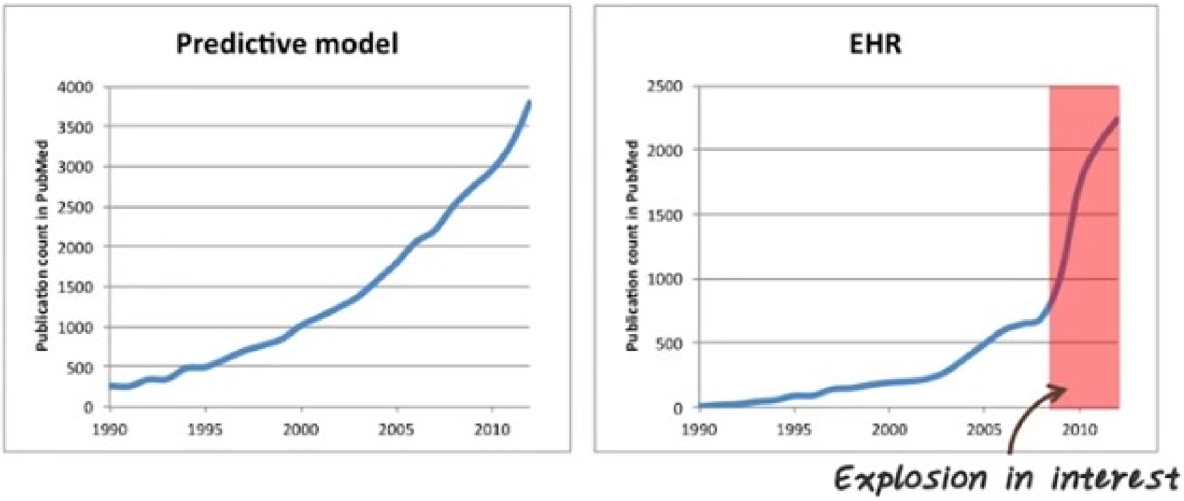
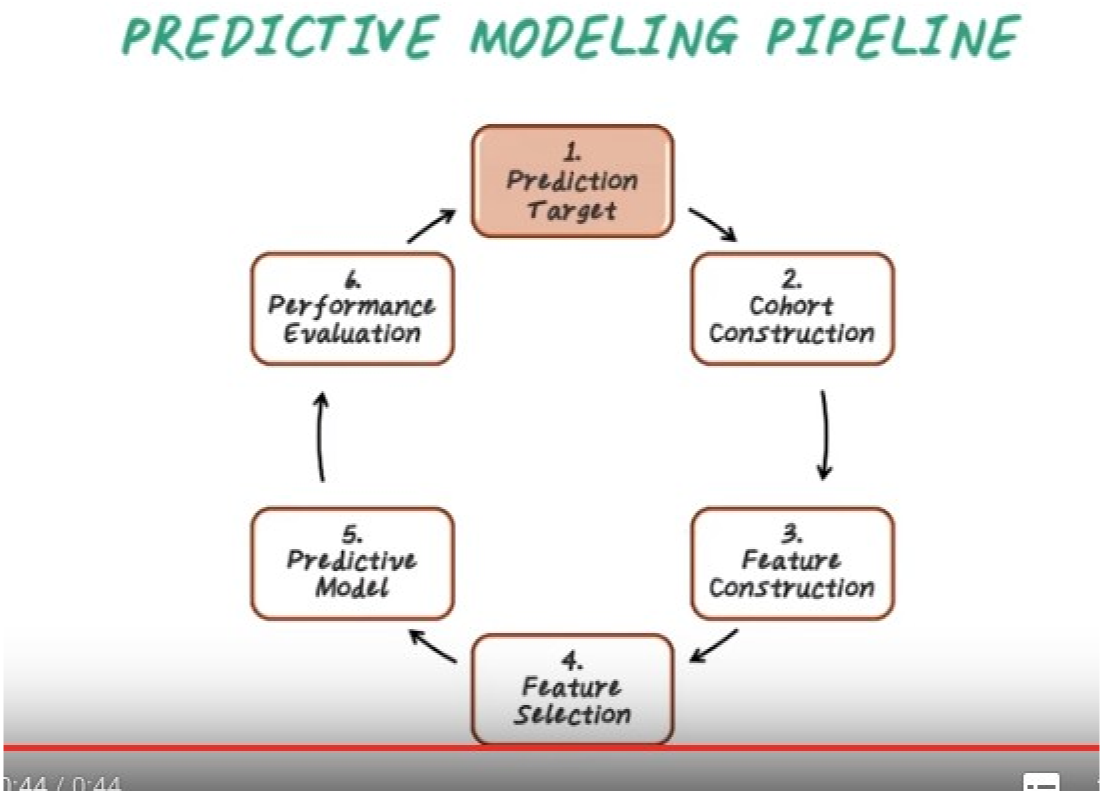
1. This lesson is about predictive modeling. What is predictive modeling? [SOUND] Not quite. Predictive modeling is a process of modeling historical data for predicting future events. For example, we want to use electronic health record that we have available to view a model of heart failure. So that we can predict patients who are at risk of developing heart failure sooner. The key goal we want to answer in this lessons is how do we develop a good predictive model using electronic health record quickly?

PREDICTIVE MODELING
VS.
ELECTRONIC HEALTH RECORDS (EHR)

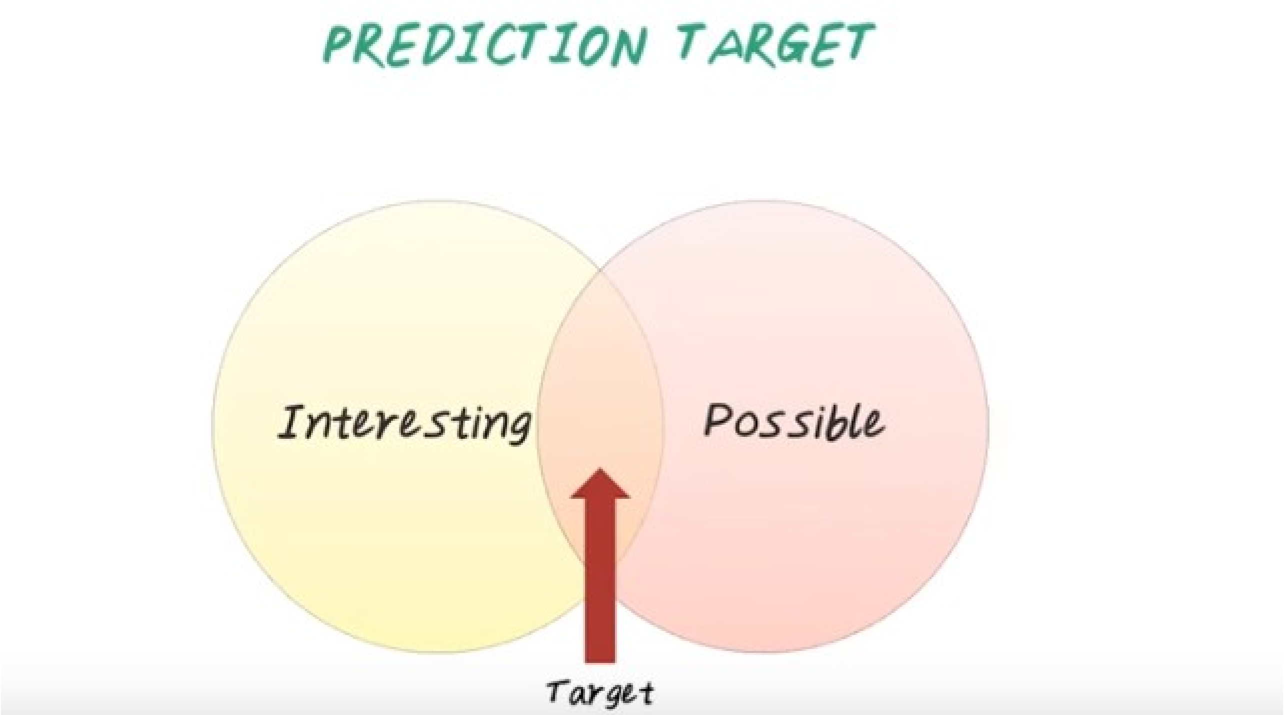


2. In this lesson we'll focus on describing how to perform predictive modeling using electronic health records, or EHR. To demonstrate the importance of predictive modeling and EHR, here we're showing the number of publications with the keyword predictive model over times and the number of publications with the keyword EHR over time. Especially in the past few years, there is an explosion of interest in EHR as EHR become a major data sources for clinical predictive modeling research. Therefore, it's important to learn how to develop a good predictive model using EHR data.



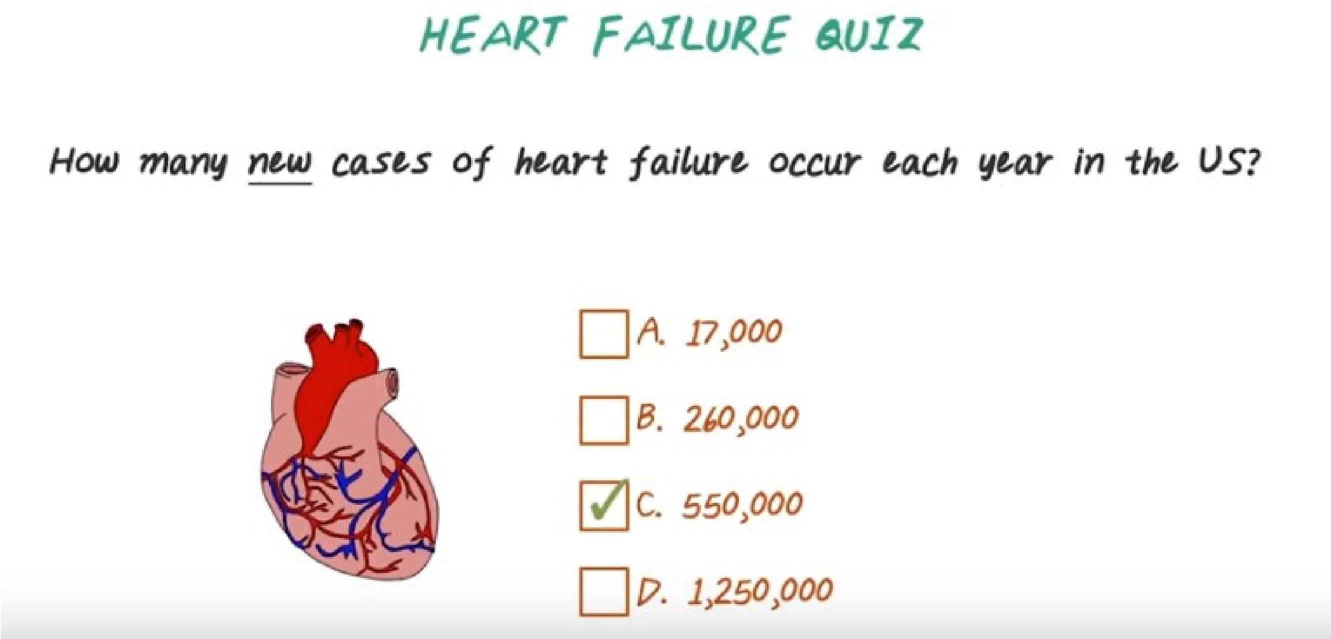
本節課後面都是在講上圖中的每一步.

3. Predictive modeling is not a single algorithm, but a computational pipeline that involves multiple steps. First, we decide the prediction target, for example, whether a patient will develop heart failure in the next few years. Second, we construct the cohort(同伴；共犯) of relevant patients for the study. Third, we define all the potentially relevant features for the study. Fourth, we select which features are actually relevant for predicting the target. Fifth, we compute the predictive model, and sixth, we evaluate the predictive model. Then we iterate this process several times until we are satisfied with the resulting model. Now let's start with prediction target.



PREDICTION TARGET

4. There are often many targets that an investigator want to predict using the data they have. However, only a subset of them are possible. So we should choose the prediction target that addresses the primary question that is both interesting to the investigator and possible to be answered using the data. For this lessons, let's focus on predicting the onset of heart failure, which is an interesting and potentially possible target.



HEART FAILURE QUIZ

How many new cases of heart failure occur each year in the US?

□ A. 17,000

□ B. 260,000

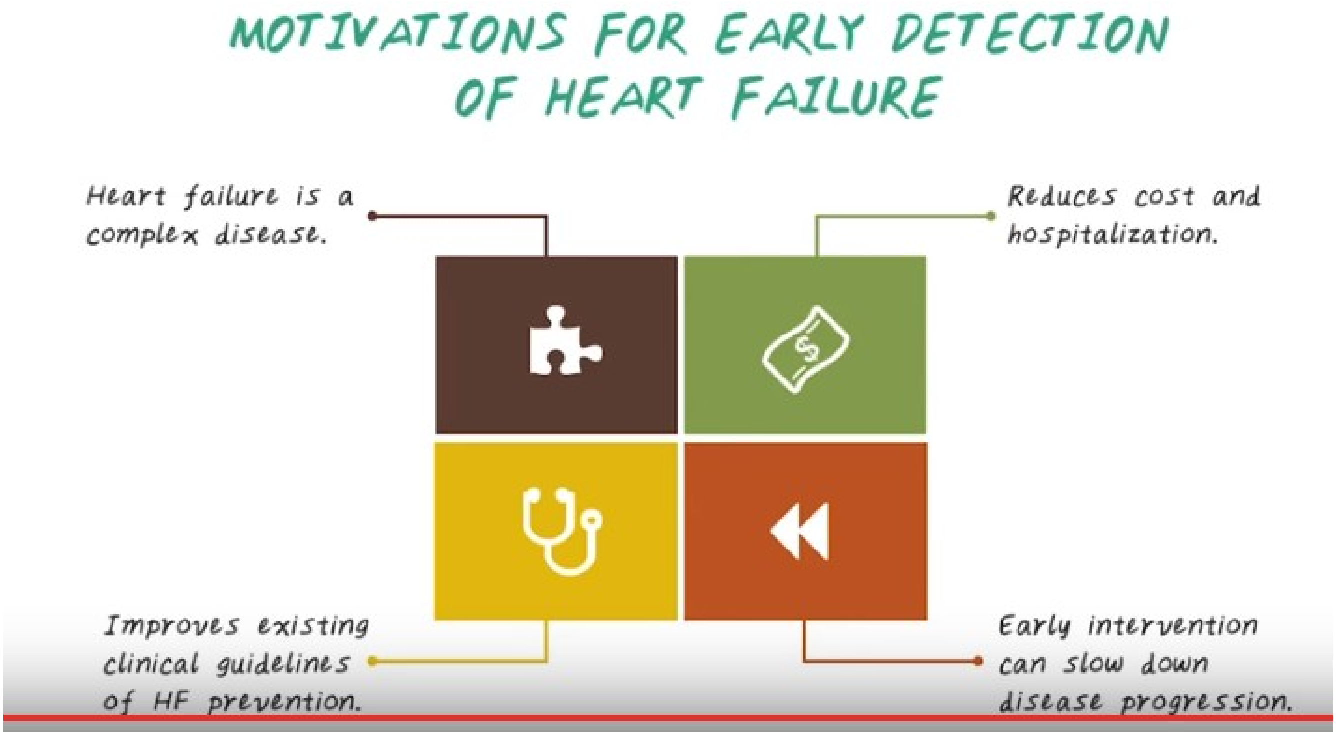
☑ C. 550,000

□ D. 1,250,000

5. Here's a quiz question on heart failure. Make a guess. How many new cases of heart failure patients

occurred each year in the U.S.? Is it a, 17,000 patients? Or b, 260,000 patients? Or c, 550,000 patients? Or d, 1,250,000 patients?

6. The correct answer is 550 thousand patients, which is a huge health care problem.



MOTIVATIONS FOR EARLY DETECTION OF HEART FAILURE

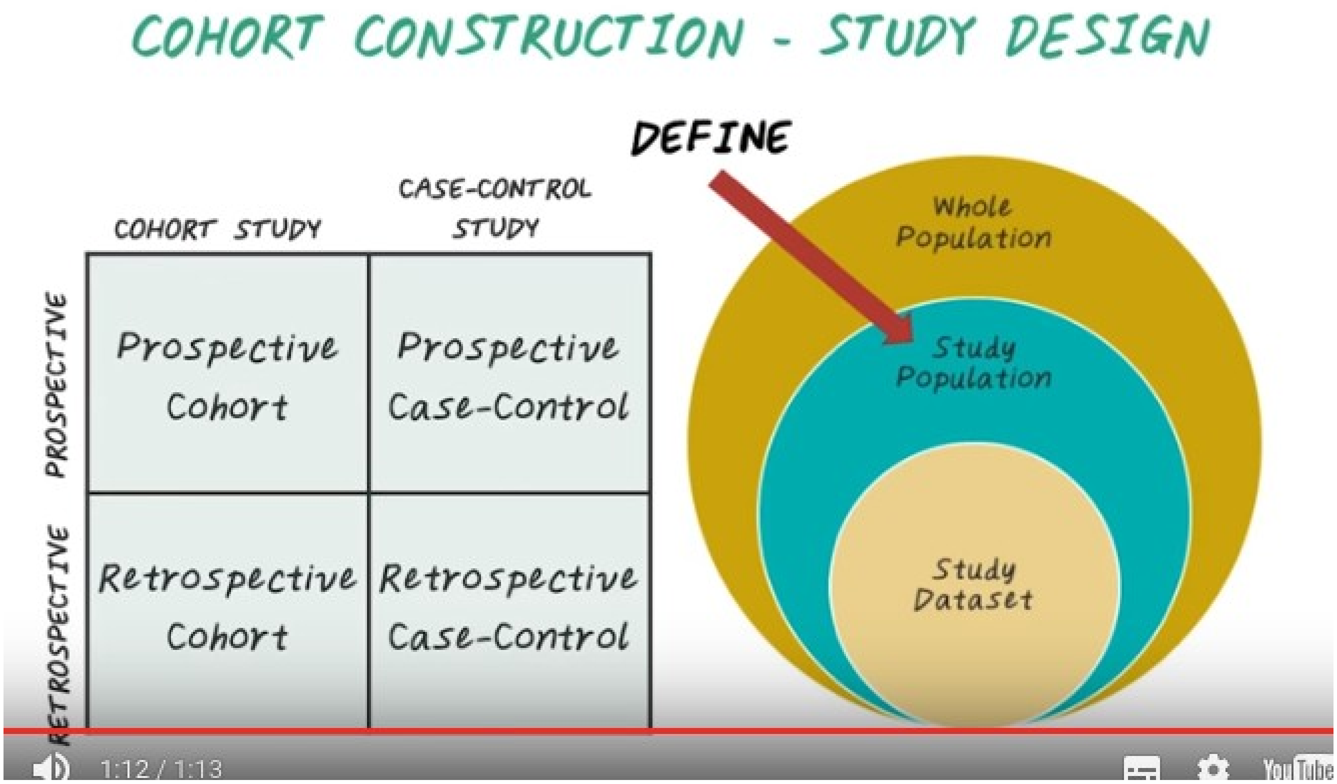
Heart failure is a complex disease.

Reduces cost and hospitalization.

Improves existing clinical guidelines of HF prevention.

Early intervention can slow down disease progression.

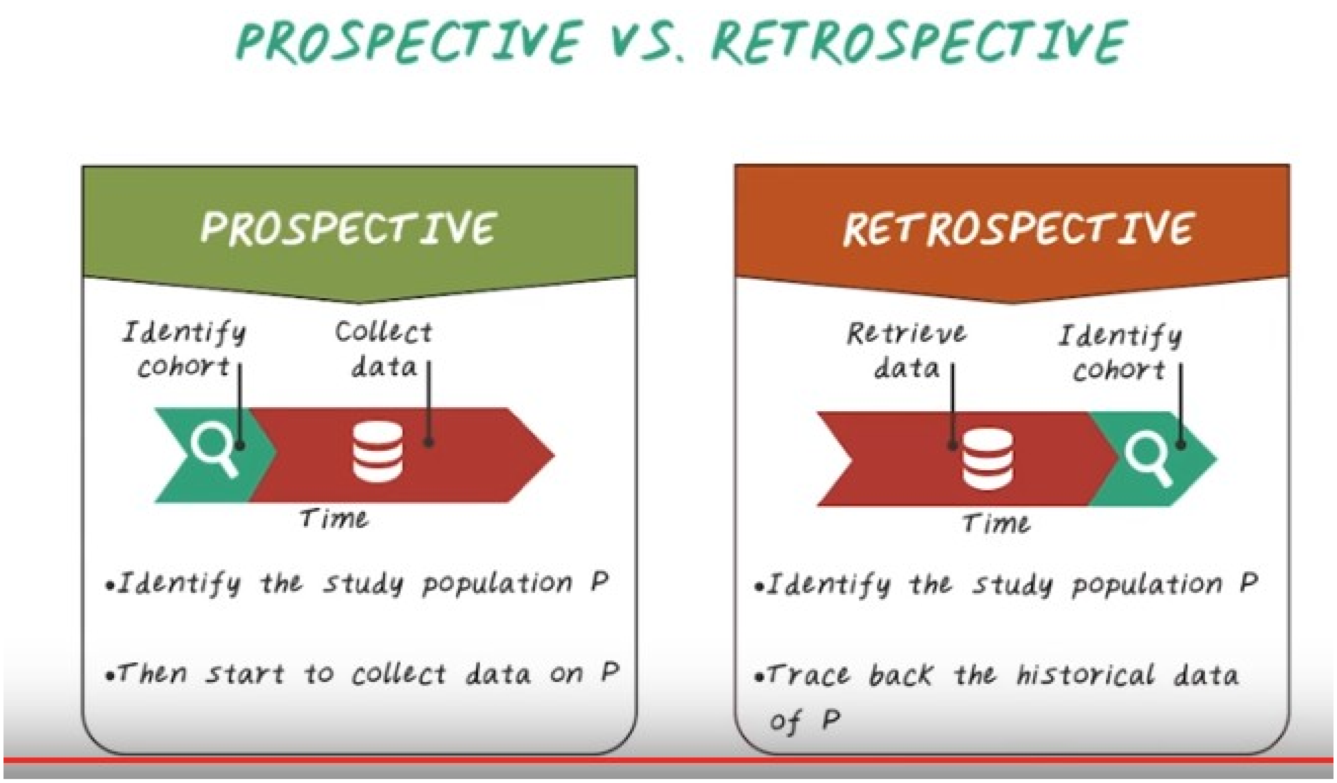
7. So we want to develop a predictive model for heart failure. But what are the motivations for early detection of heart failure? First, heart failure is a complex disease. There is no widely accepted characterization and definition of heart failure. Probably because the complexity of the syndrome. It has many potential ideologies, diverse clinical features, and numerous clinical subsets. If we can detect heart failure earlier, we can potentially reduce the cost of hospitalization associated with heart failure. We can also potentially introduce new early intervention to try to slow down the progression of heart failure, improve the quality of life, and reduce mortality. In the long term we can improve existing clinical guidelines for heart failure prevention. So in this class we'll show you how to develop a predictive model for predicting heart failure earlier.



COHORT CONSTRUCTION - STUDY DESIGN

	COHORT STUDY	CASE-CONTROL STUDY
PROSPECTIVE	Prospective Cohort	Prospective Case-Control
RETROSPECTIVE	Retrospective Cohort	Retrospective Case-Control

8. So far, we're talked about how to define the prediction target. Next, we introduce the Cohort Construction step. Cohort construction is about defining the study population. For a given prediction target, there are only a subset of patient that are relevant among the whole patient-population. And they are the Study Population. Be aware, often may not be possible to obtain data from everybody in the study population. As a result, the data set we studied is only a subset of those Study Population. So the question is, how do we define the Study Population? There are two different axes to be considered. One the vertical axis, we have prospectives study versus retrospective study. On the horizontal axis, we have cohort study verses case-control study. Depending on the combinations, we have four different options. Perspective Cohort study, Perspective Case-Control study, Retrospective Cohort study, and Retrospective Case-Control study. Now let's let's look at this two axis in more details.



PROSPECTIVE VS. RETROSPECTIVE

PROSPECTIVE

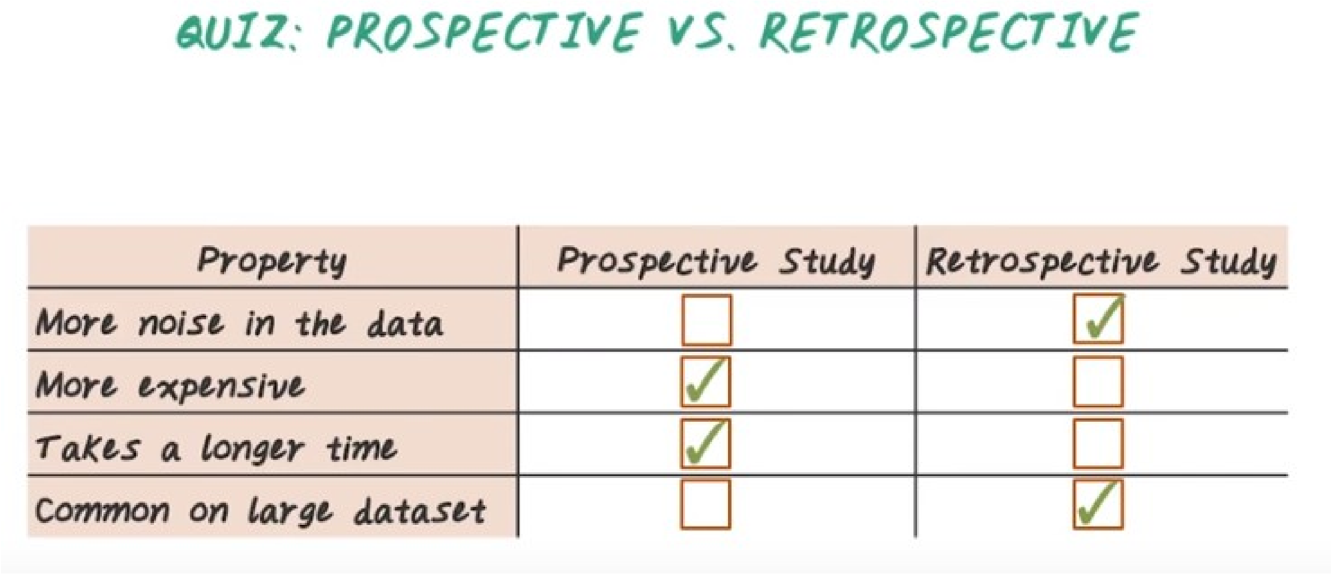
- Identify the study population P
- Then start to collect data on P

RETROSPECTIVE

- Identify the study population P
- Trace back the historical data of P

9. Now, let's talk about prospective versus retrospective studies. In a prospective study, we first identify the cohort of patients, then decide what information to collect and how to collect them. Then start the data collection. In contrast, in a retrospective study, we first identify the patient cohort from existing data. For example past electronic health records of patients, then retrieve all the data about the cohort. So, in prospective study, we identify the cohort and collect the data from scratch. But in the retrospective study, the data set already exists. We just need to identify the right subset and retrieve them.

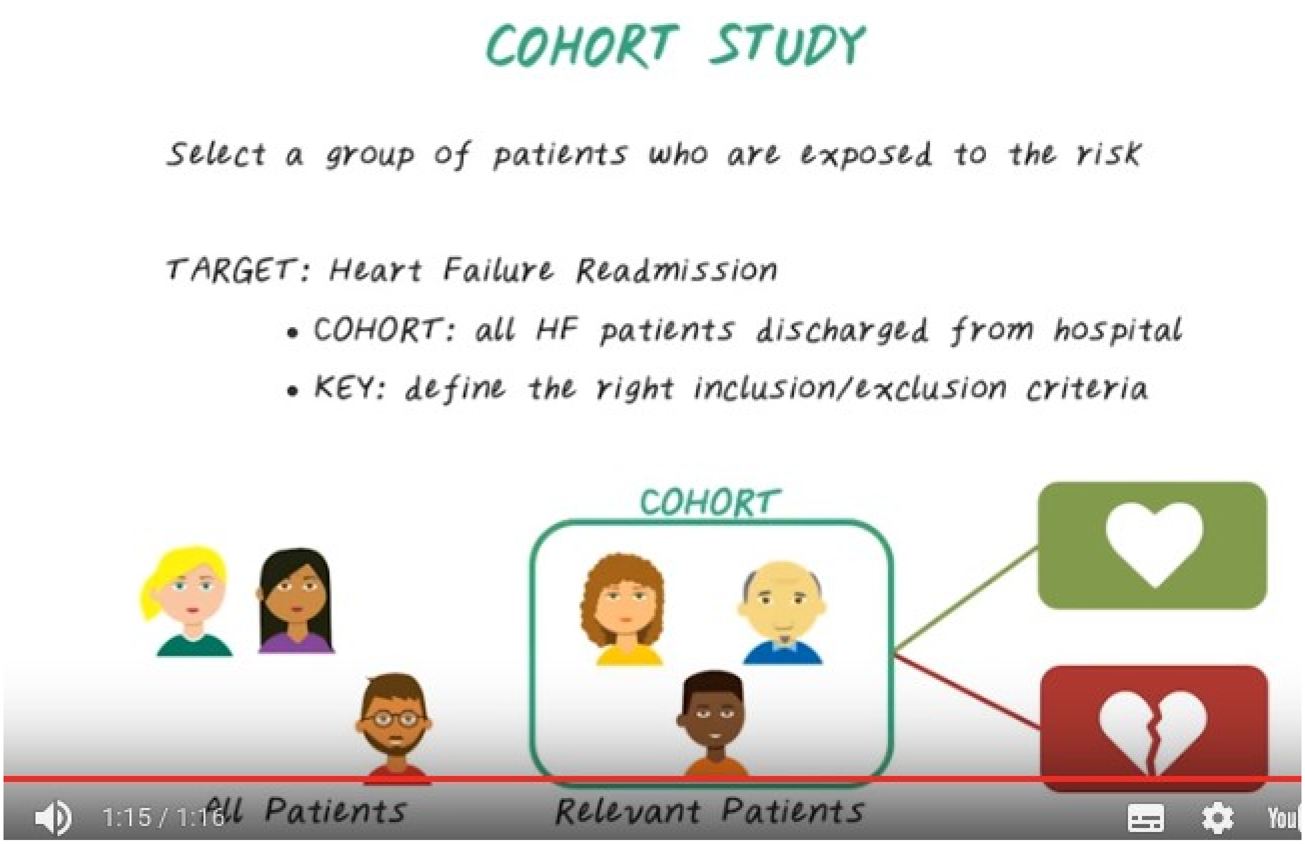
10. Here's a quiz question to compare perspective and retrospective studies. Each row represents a particular property. Pick the study that has the corresponding property. More specifically, which one has more noise in their data? Which one is more expensive to conduct? Which one takes a longer time to conduct, and which one is more commonly done on larger dataset?



QUIZ: PROSPECTIVE VS. RETROSPECTIVE

Property	Prospective Study	Retrospective Study
More noise in the data	□	☑
More expensive	☑	□
Takes a longer time	☑	□
Common on large dataset	□	☑

11. Here are the answers. So, retrospective study often work on data with more noise, because data are often created for other purpose, not research. In contrast, prospective study, because you design a dataset collection process specifically for this research, the quality of the data is often higher. As a result, less noise. Prospective study is often more expensive and takes longer time to conduct because the data has to be collected from scratch. Finally because the cost and time constraints, the size of the data set prospective study used is often smaller and limited. On the other hand, retrospective study deal with historical data. It often can work with much larger data set.



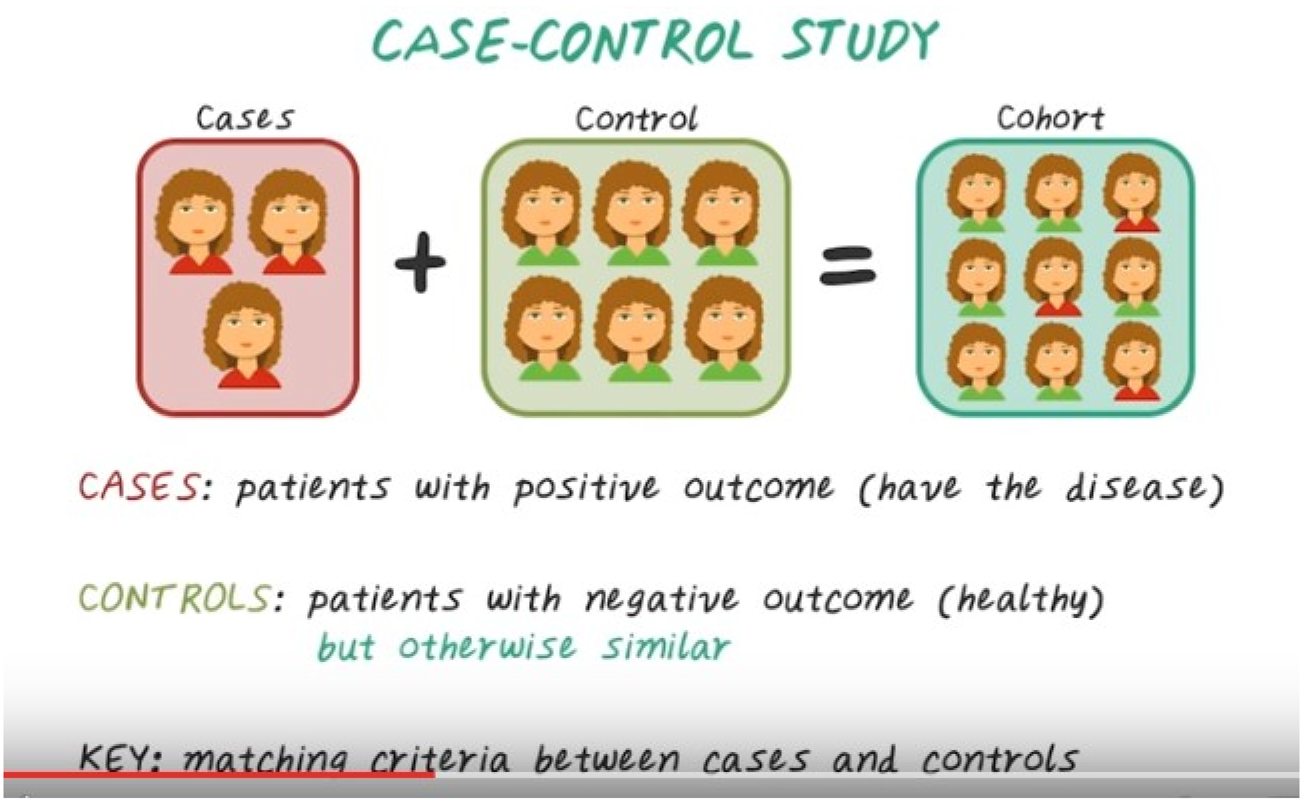
COHORT STUDY

Select a group of patients who are exposed to the risk

TARGET: Heart Failure Readmission

- COHORT: all HF patients discharged from hospital
- KEY: define the right inclusion/exclusion criteria

12. Next, let's talk about COHORT study. In a COHORT study, the goal is to select a group of patients who are exposed to a particular risk, for example, if we want to be in a predictive model for predicting heart failure readmission. Here heart failure readmission means, heart failure patient after discharged from the hospital, comes back again to the hospital due to heart failure. In this case, the COHORT should contain all the heart failure patients who discharged from the hospital, because they can potentially be readmitted after discharge. The key in COHORT study is to define the right inclusion and exclusion criteria to figure out what patient to include. Here's a visual illustration. We start with all patients, then we try to identify the relevant patient for a particular risk, for example, these three patients are relevant for this risk, such as heart failure readmission. Then we want to build a model to predict the target risk. In this case, the COHORT contains both positive and negative examples, for example, patient with heart failure readmission and patient without heart failure readmission. All this relevant patient is a COHORT in this study.



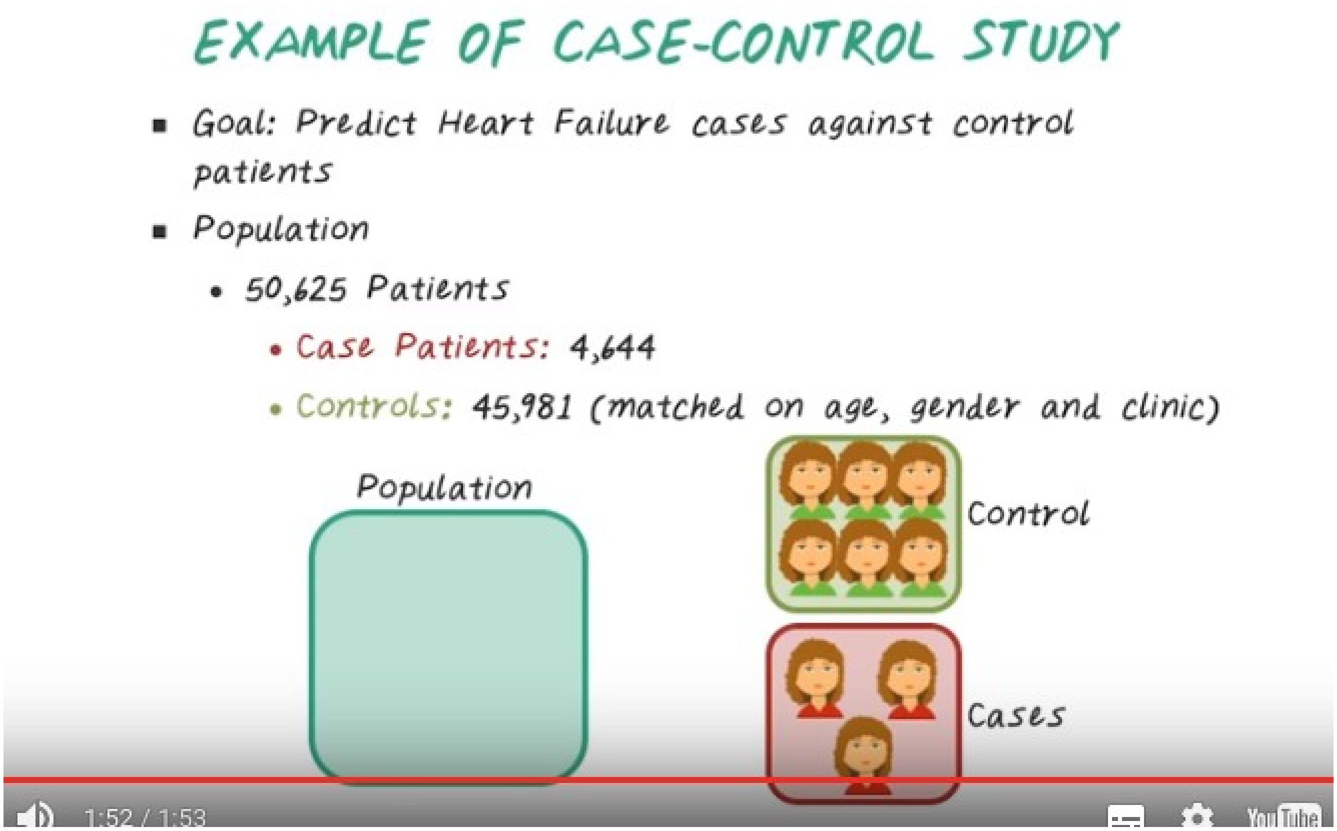
CASE-CONTROL STUDY

CASES: patients with positive outcome (have the disease)

CONTROLS: patients with negative outcome (healthy) but otherwise similar

KEY: matching criteria between cases and controls

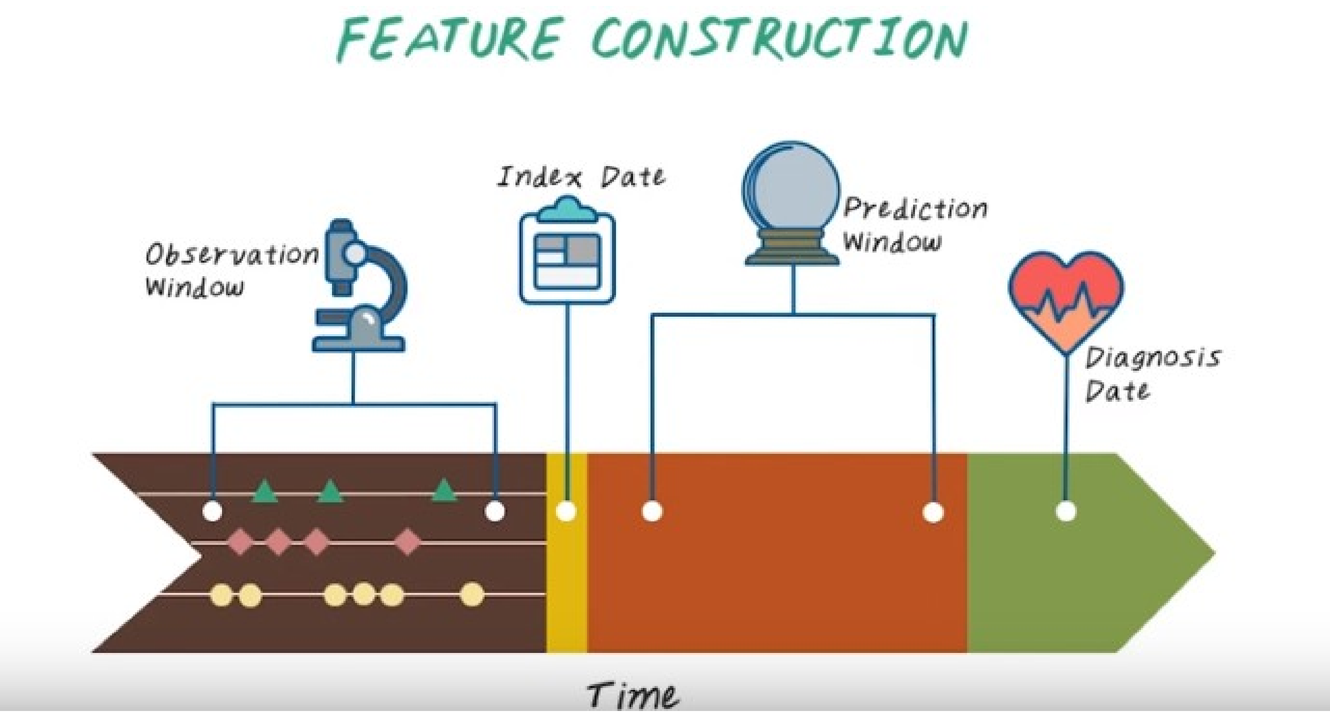
13. The other common study design is case-control study. In this design, we try to identify two sets of patient, namely cases and controls. And we put them together to construct the cohort. Cases are patients with positive outcome.



EXAMPLE OF CASE-CONTROL STUDY

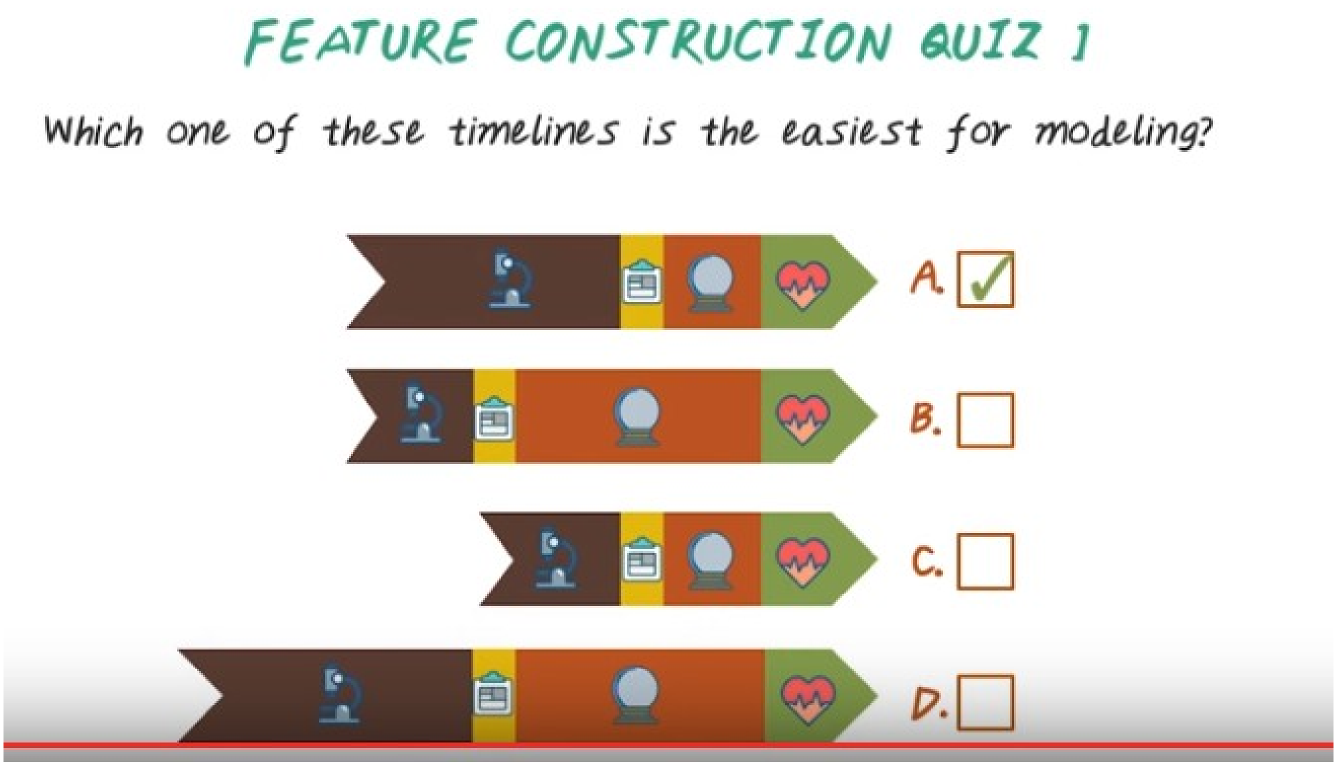
- Goal: Predict Heart Failure cases against control patients
- Population
  - 50,625 Patients
    - Case Patients: 4,644
    - Controls: 45,981 (matched on age, gender and clinic)

For example, the patient who develop the disease. Controls are the patients with negative outcomes. That is, they're healthy patients, but otherwise similar to the cases. For example, they can have the same age, gender, and visit the same clinics. And the key here is to develop the matching criteria between cases and controls. For example, we want you to predict a model of heart failure. Then we identify a study population of over 50,000 patients, and we have 4,644 case patients. Those are the patients who developed heart failure. And we matched them against a set of control patients on age, gender, and clinics. And we end up with 45,000 control patients. Notice that in this study, we have a lot fewer cases with heart failures than the controls without heart failures. This is pretty typical because in a real world scenario, patients with the specific disease conditions are often harder to obtain, while there are a lot more patients without that disease are available to serve as a control. 講得好: To summarize, in a case-control study, we first identify the cases, then try to match them to a set of control patients. In a cohort study, we'll identify all the patients who are exposed to the risk and the matching criterias are not involved.



14. So far we talked about how to define the prediction target, how to construct the patient cohort, next we introduce feature construction step. The goal of feature construction is to construct all potentially relevant features about patients in order to predict the target outcome. Next, we introduce a few key concepts that are related to feature construction. First, the raw patient data arriving as event sequences over time. Diagnosis date is the date that the target outcome happened. In the heart failure predictive modeling example, each patient is diagnosed with heart failure on this date. Since control patient does not have heart failure diagnosis, in theory, we can use any days from control patient as the diagnosis date. But commonly we choose to use the heart failure diagnosis date of the matching case patient as diagnosis date for the corresponding control. Before the diagnosis day, we have a time window, called prediction window. Before the predication window, we have the index day at which we want to use the learn predicted model to make a predication about the target outcome. Before the index day, we have another time window called observation window. We use all the patient information happening during this observation window to construct features. There are many different ways to construct features. For instance, we can count the number of times an event happens. For example, if type two diabetes code happened three times during this observation window, the corresponding feature for type two diabetes equals three. Or sometimes we can take average of the even value. For example, if patient has two HBA1C measures during observation window, we can take the average of this two measurement as a feature for HBA1C. The length of prediction window and observation window are two important parameters that going to impact the model performance. Next, we illustrate their impact using some examples.

15. Here's Chris to help us understand the impact of prediction window and observation window. Which of the following timelines is easiest for modeling? Is it A, large observation window and small prediction window? Or B, small observation window and large prediction window? Or C, small observation window and small prediction window. Or D, large observation window and large prediction window.



FEATURE CONSTRUCTION QUIZ 1

Which one of these timelines is the easiest for modeling?

A. ☑

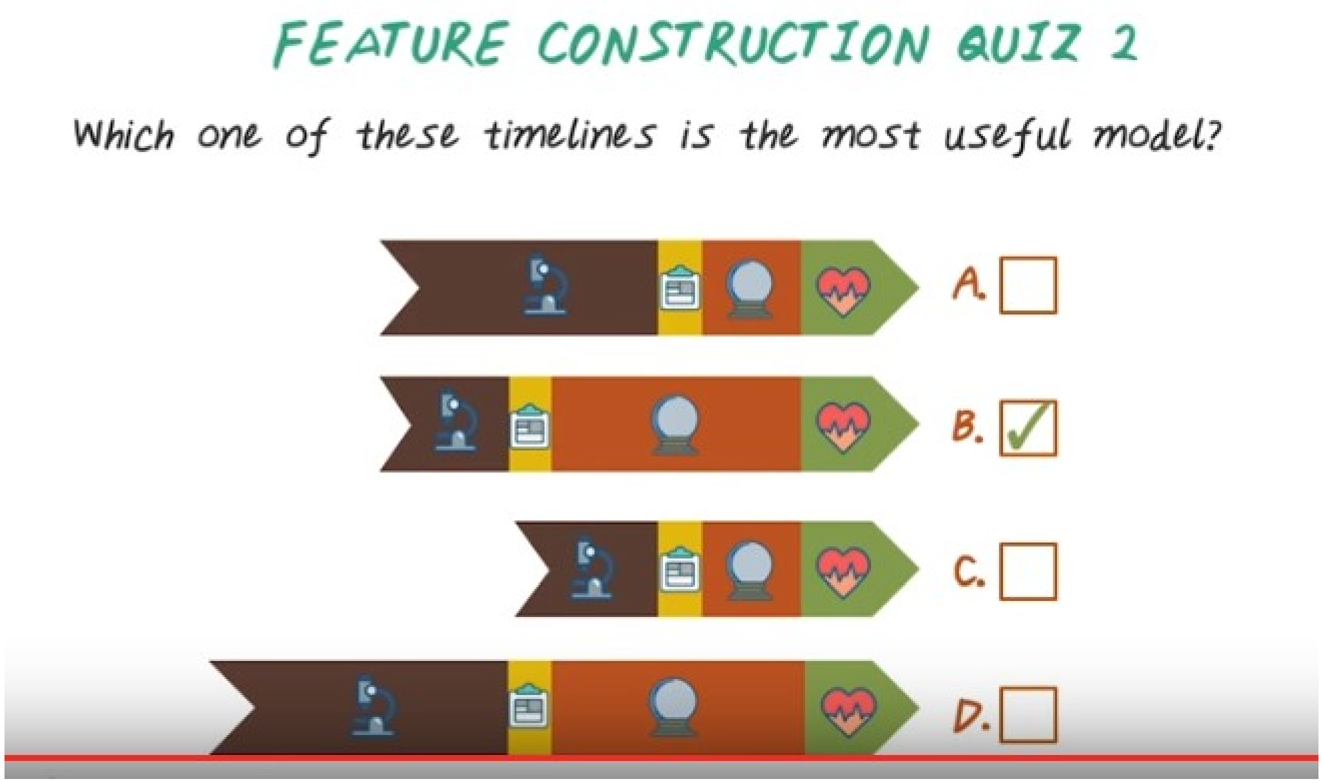
B. □

C. □

D. □

16. The answer is A, large observation window and small prediction window. Because it is often easier to predict event in the near future, that is, small prediction window. On the other hand, large observation window means more information to be used to construct features, which is often better since we can model patient better with more data. Therefore, large observation window and small prediction window is easiest for modeling.

17. Here is another quiz. Which of these timeline is the most useful model? Large observation window, small prediction window. Or small observation window, large prediction window. Or small observation window and small prediction window. Or large observation window and large prediction window.



FEATURE CONSTRUCTION QUIZ 2

Which one of these timelines is the most useful model?

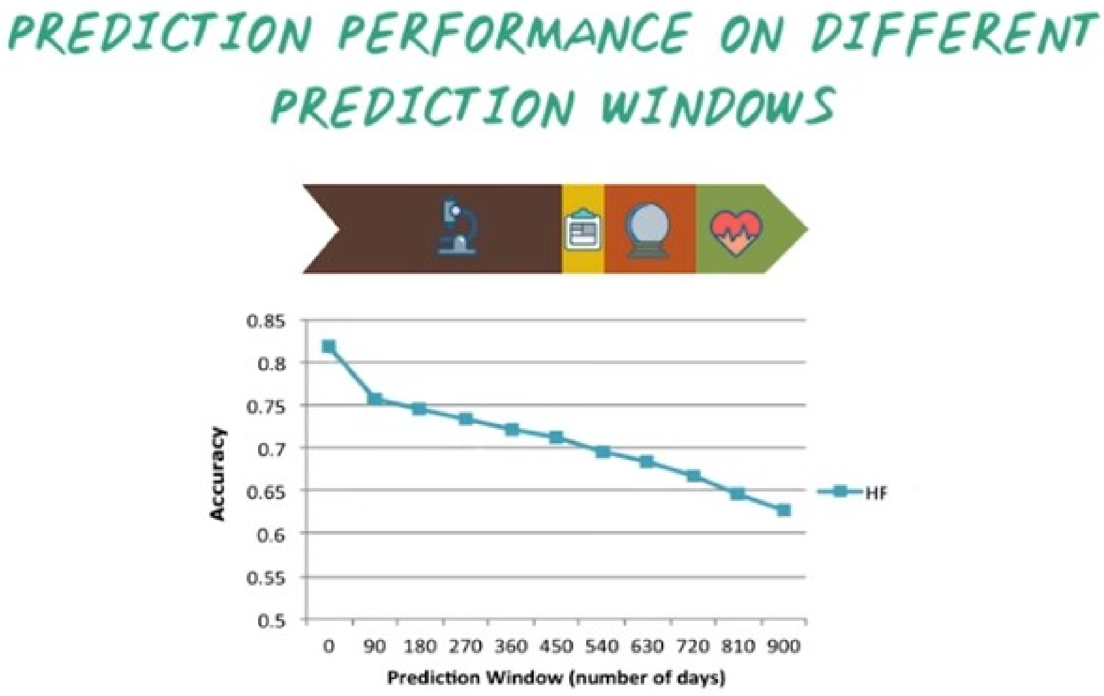
A. □

B. ☑

C. □

D. □

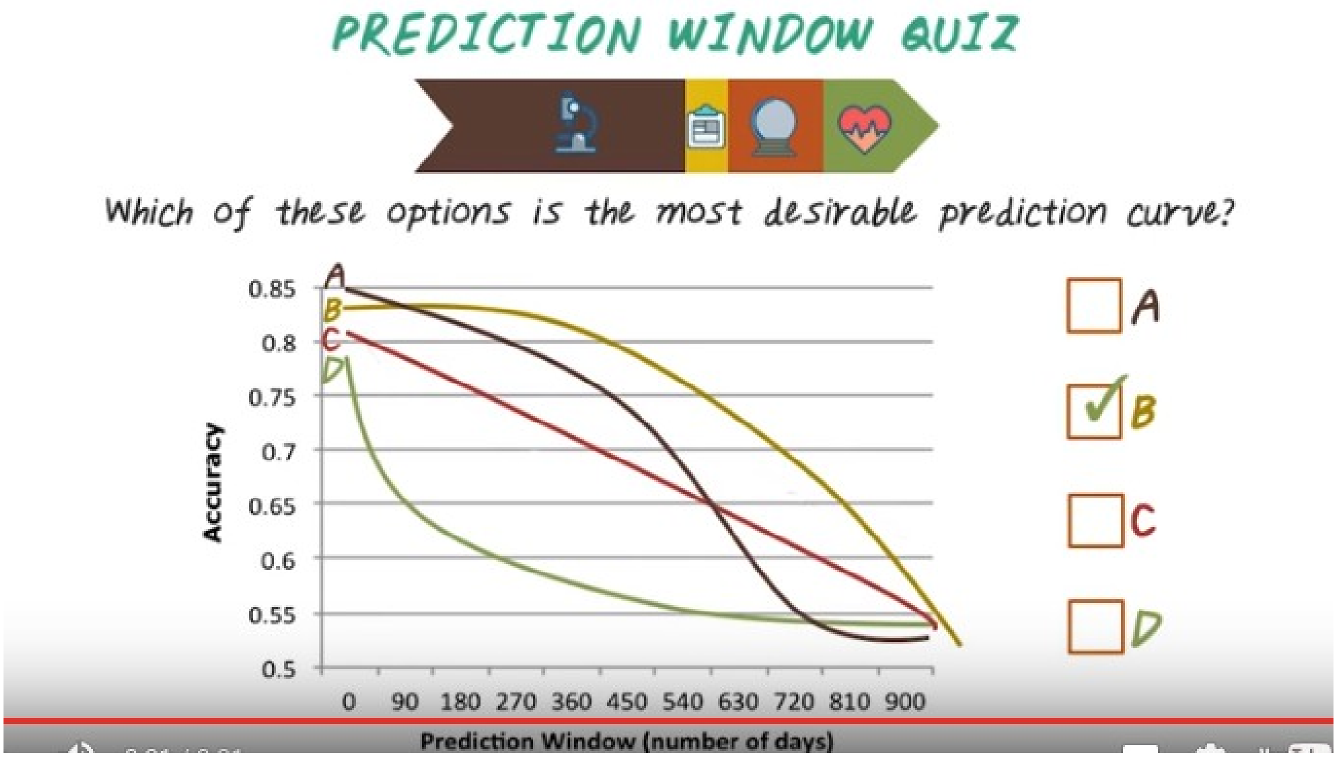
18. The answer is B. Small observation window, large prediction window. In this idea situation, if we can construct a good model, we want to predict far into future. Therefore, large prediction window, without much data about the patient. Therefore small observation window. That's why B reflects idealistic timeline. If we compute a model in this setting, this will be the most useful model. However, the setting is often difficult to model, therefore unrealistic.



PREDICTION PERFORMANCE ON DIFFERENT PREDICTION WINDOWS

19. Here's another example illustrating the impact of prediction window. In this chart, the y axis is the accuracy of the model, the higher the better. The x axis is the size of the prediction window, which varies from zero days to 900 days. We can see the accuracy of the model drops as we increase the prediction window. Because, it's easier to predict things happen in near future, than things happen far into the future.

20. Here's another quiz question on prediction window. Which of the following options is the most desirable prediction curve? Is it A, or B, or C, or D?



PREDICTION WINDOW QUIZ

Which of these options is the most desirable prediction curve?

□ A

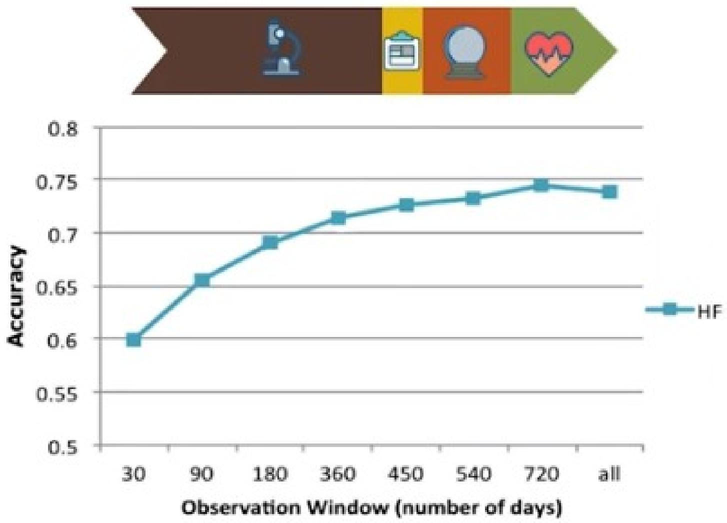
☑ B

□ C

□ D

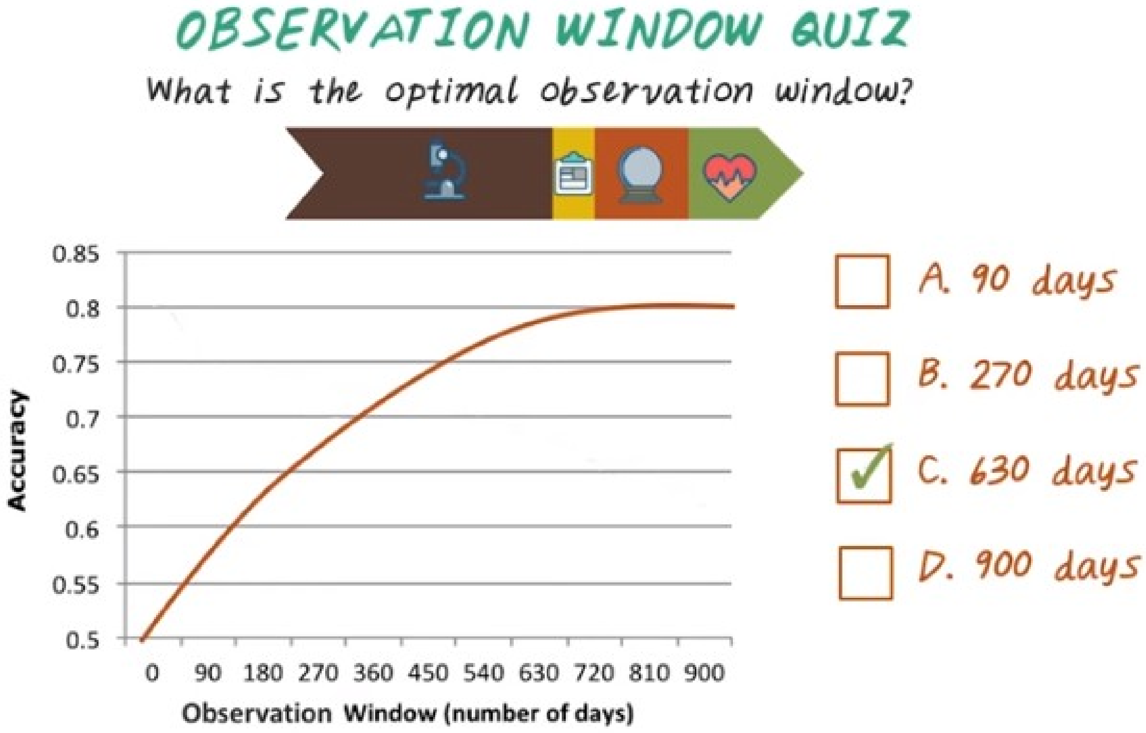
21. The answer is B because we can predict accurately for fairly long periods of times up to 450 days of prediction window. While the performance of all the other model drops fairly quickly as the prediction window increases. You may notice that A has the maximum accuracy at the beginning. However, as the prediction window increases, the performance of A drops quite quickly. So it's not that of useful model for predicting long term things.

PREDICTION PERFORMANCE ON DIFFERENT OBSERVATION WINDOWS



22. Now let's consider the performance of different observation windows. Typically, as the observation window increases, the performance improves because you know more about the patient as the observation window increases.

23. Here's the quiz on observation window. Given the performance curve, when we vary the observation window like this. What is the optimal observation window we should choose? Is it A, 90 days, or 270 days, or C, 630 days, or D, 900 days?



OBSERVATION WINDOW QUIZ

What is the optimal observation window?

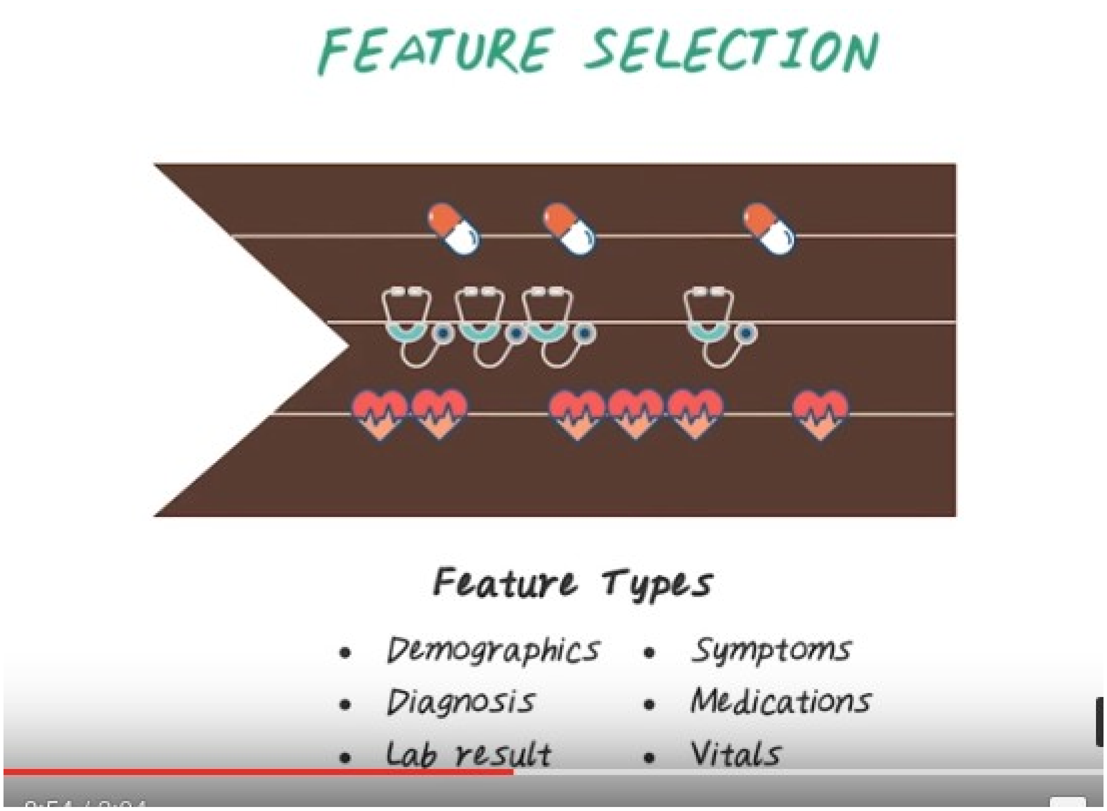
□ A. 90 days

□ B. 270 days

☑ C. 630 days

□ D. 900 days

24. The answer is C, because the model performance plateaued after 630 days. It indicates a diminishing return as we go further beyond that point. Therefore 630 days is a good choice. So you may wonder, choosing 900 days may also be a good choice, but it's a trade-off between how long is the observation window, and how many patients have that much data. So, if you chose 900 days as observation window for patients who do not have enough data up to 900 days, they will be excluded from the study.

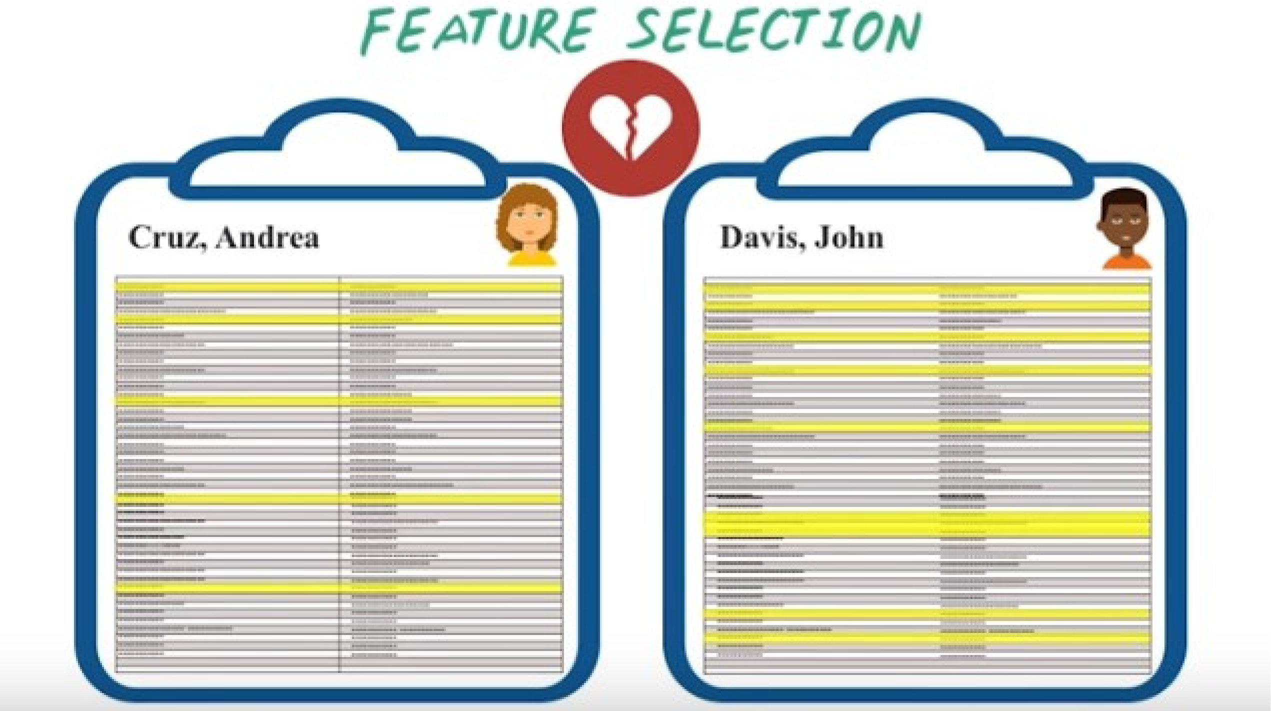


FEATURE SELECTION

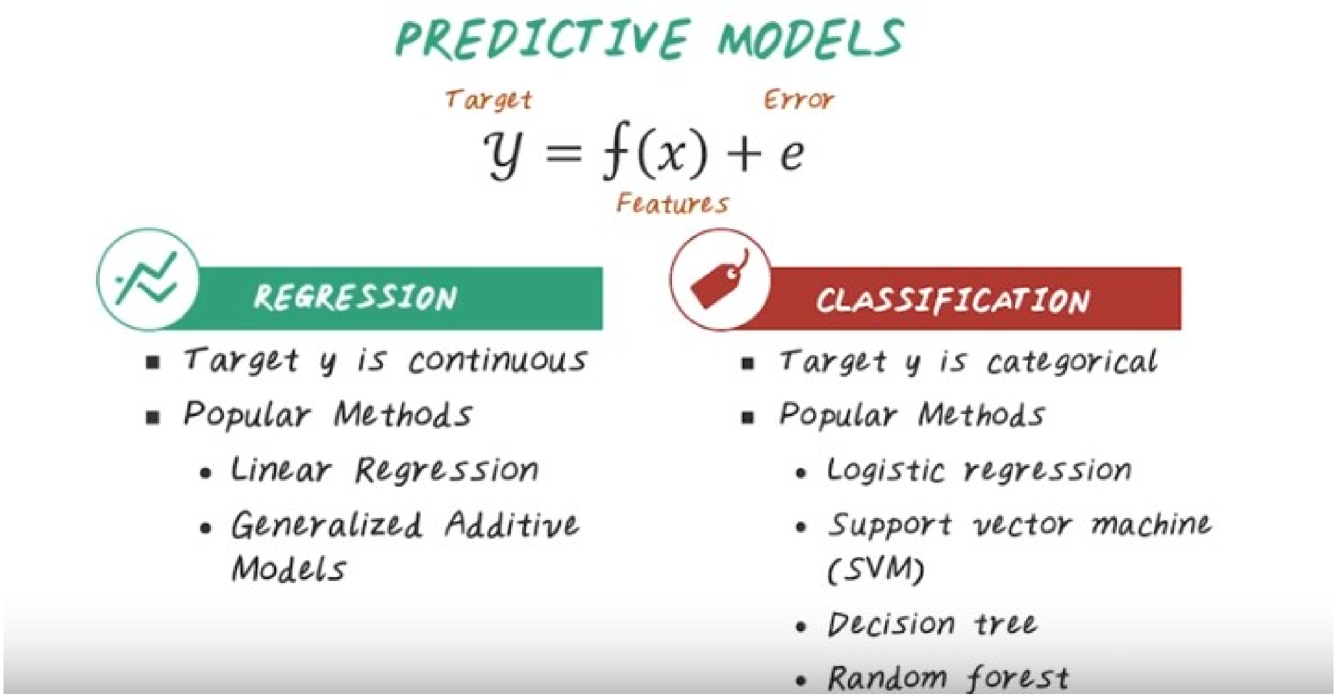
Feature Types

- Demographics
- Diagnosis
- Lab result
- Symptoms
- Medications
- Vitals

25. We covered the first three steps. Next we'll talk about the feature selection step. In the feature selection step we have talked about how construct features using patient event sequences from the HRC data. In particular, we construct feature from raw data in the observation window. If we look closely at the observation window, we see event sequence data, which are corresponding to different types of clinical events. For example, diagnosis, symptoms, medications, patient demographics, lab results, and vital signs. We can construct features from all those events. However, not all the events are relevant for predicting a specific target. The goal of feature selection is to find the truly predictive features to be included in the model.



For example, here are two patient charts. We can see some features such as demographics, including age, race, and gender, and vital signs such as blood pressures, and diagnosis such as diabetes and hypertension. However, in reality this patient chart is not that simple. In fact, we can construct a long list of features over 20,000 features, from a typical EHR data set. Not all of this are relevant for predicting a target. We need to select the ones that are relevant to the target condition. For example, if we want to predict heart failure, maybe those yellow features are relevant. However, for a different condition, such as, diabetes. Maybe those purple features are relevant. The goal of features selections is to identify those predictive features. Giving a specific target condition.



PREDICTIVE MODELS

Target Error

$$y = f(x) + e$$

Features

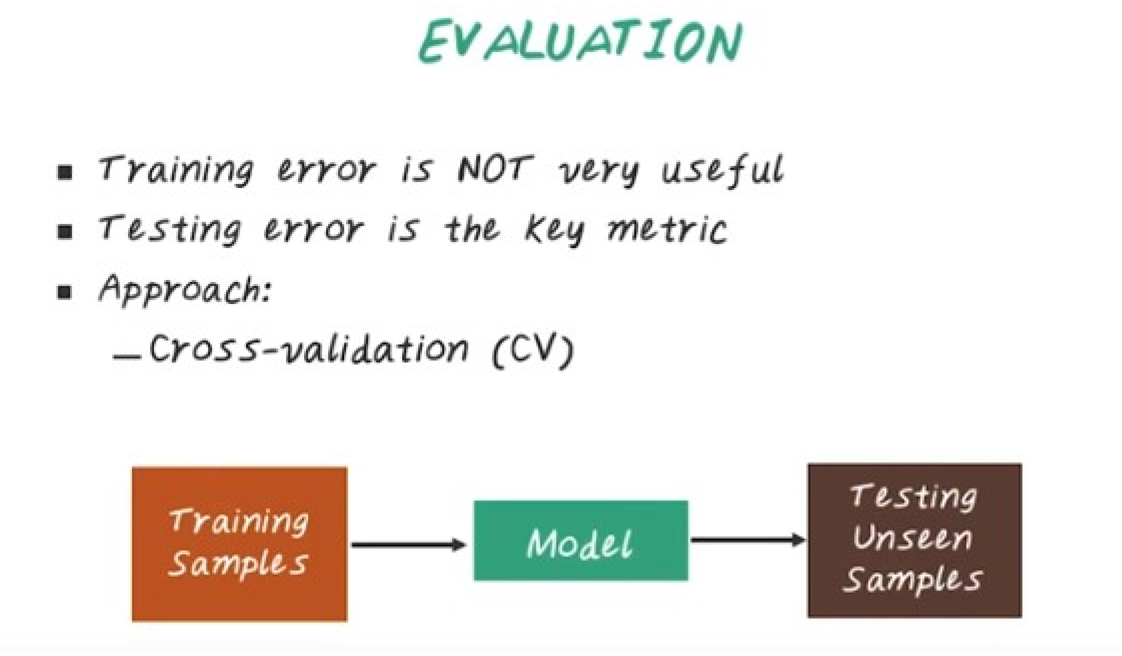
REGRESSION

- Target y is continuous
- Popular Methods
  - Linear Regression
  - Generalized Additive Models

CLASSIFICATION

- Target y is categorical
- Popular Methods
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree
  - Random forest

26. So far we've figured out what we want to predict and who we will use to make the prediction and what feature is to be used in this prediction. And next let's see how we make the prediction (Predictive Model 這一步). Predictive model is the function that maps the input features of the patient to the output target. For example, if we know a patient's past diagnosis, medication, and lab result, if we also know this function, then we can assess how likely the patient will have heart failure. Depending on the value of the target, the model can be either regression problems or a classification problem. In regression problem, the target is continuous. For example, if we want to predict the cost that a patient will incur to the healthcare systems, then it's a regression problem, and y is the cost in dollars. And the popular method includes linear regression and generalized additive model. And if the target is categorical, for example, whether the patient has heart failure or not, then it's a classification problem. Popular method include logistic regression, support vector machine, decision tree, and random forest. You may have learned all those methods in other courses such as machine learning. And in this course, we'll utilize all those methods again in the context of building a predictive model for solving healthcare problems.



EVALUATION

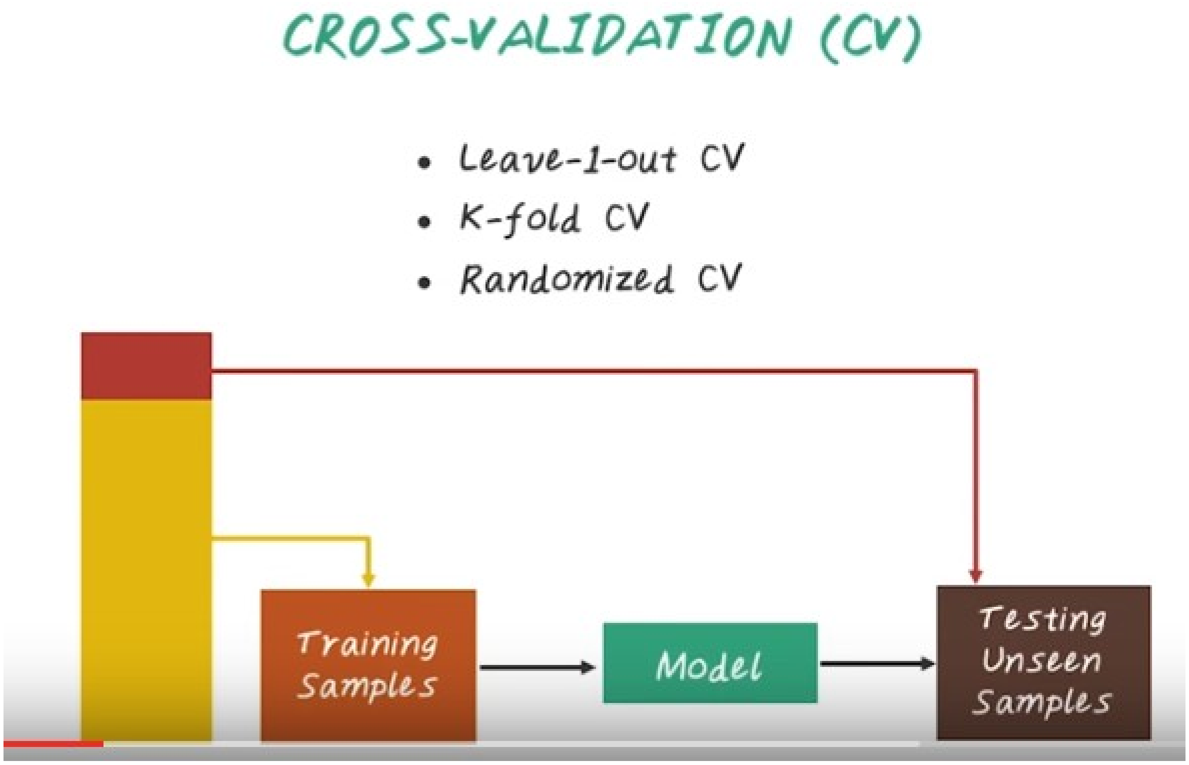
- Training error is NOT very useful
- Testing error is the key metric
- Approach:
  - Cross-validation (CV)

Training Samples

Model

Testing Unseen Samples

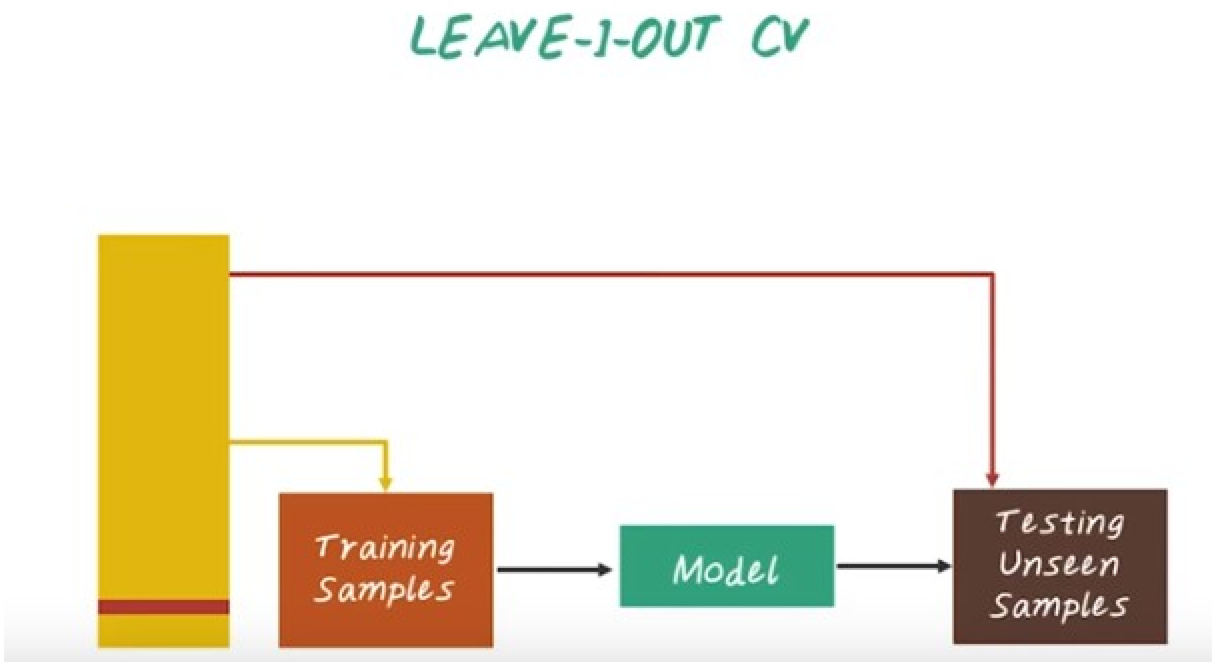
27. The final step of this pipeline is to assess how good our model is through performance evaluation. Evaluation of predictive models is one of the most crucial steps in the pipeline. The basic idea is to develop the model using some training samples, but test this train model on some other unseen samples, ideally from future data. It is important to note that the training error is not very useful, because you can very easily over fit the training data by using complex models which do not generalize well to future samples. Testing error is the key metric because it's a better approximation of the true performance of the model on future samples. The classical approach for evaluation is through cross-validation process or CV.



CROSS-VALIDATION (CV)

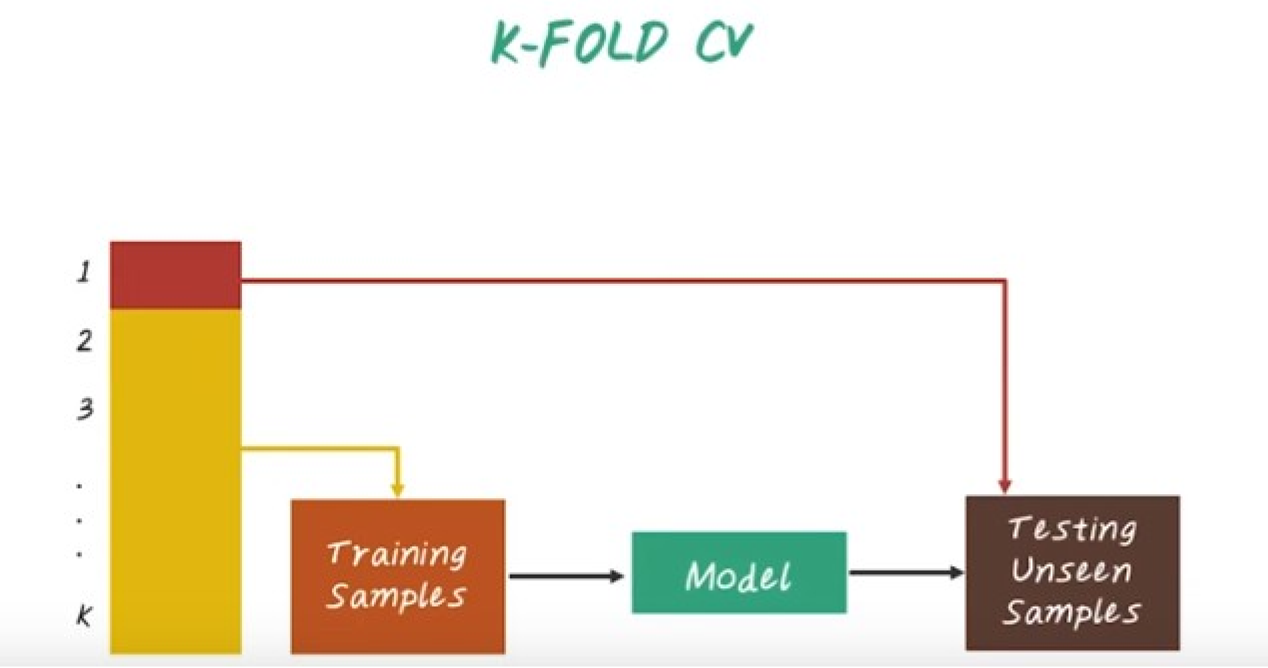
- Leave-1-out CV
- K-fold CV
- Randomized CV

28. Now, let's talk about cross-validation. The main idea behind cross-validation is to iteratively split a data set into training and validation sets. And we want to view the model on the training set, and test the model on the validation step, but do this iteratively, many times. Finally the performance matrix are aggregated across this iterations often by taking the average. There are three common messes for cross-validations namely Leave-1-out cross-validation, k-fold cross-validation, and randomized cross-validation.

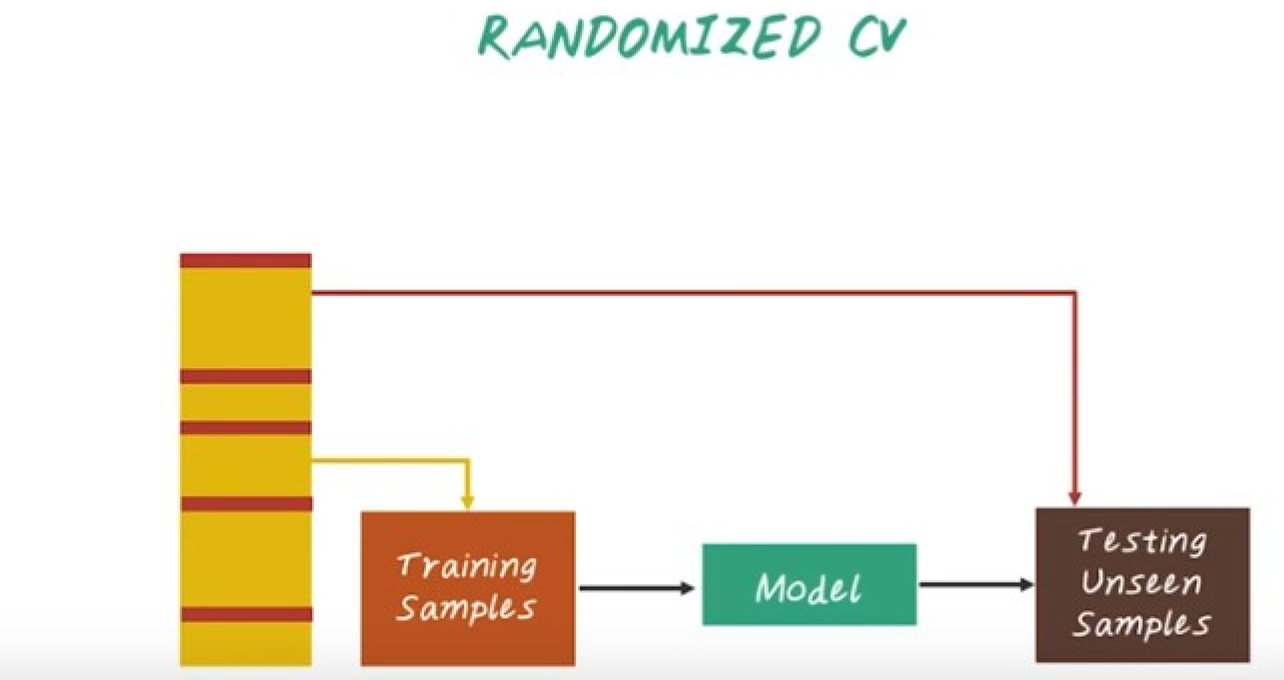


LEAVE-1-OUT CV

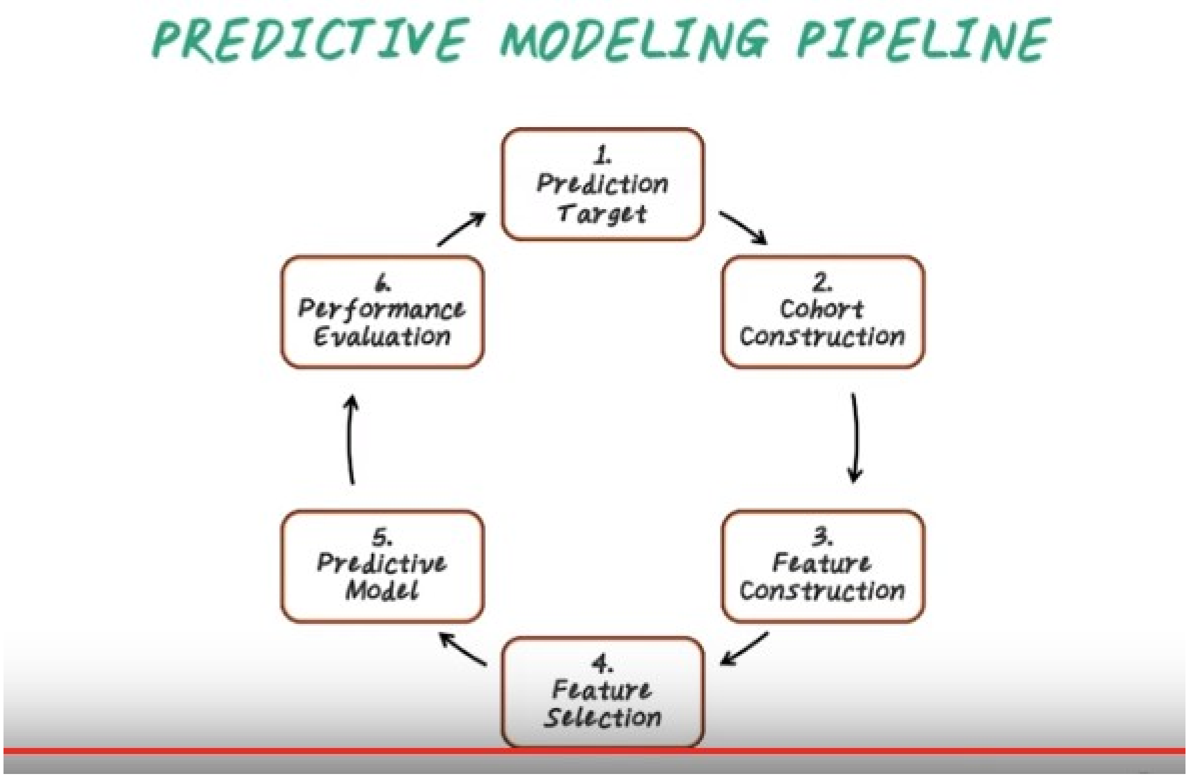
In Leave-1-Out cross validation, we take one example at at time as our validation set and use the remaining set as the training set. Then repeat this process many times, goes through the entire data set. The final performance is computed by averaging the prediction performance across all iterations.



K-Fold's cross variation is very similar to leave-1-out validation. But instead of just using one example of validation set, we have multiple examples in the validation set. More specifically, we split the entire data set into K-Folds. And we iteratively choose each fold as set, validation set and use the remaining Folds as a trimming set. For example, the Fold 1 would be used as the validation set, and the remaining fold will be used as a trimming set to view the model. Then we use a Fold 2 as the validation set, the remaining fold as the training set to be build another model. And repeat this process K times, and the final performance is the average over this four different models.



Finally, randomized cross validation will randomly split the data set into training and testing. For each such split, the model is fit to the training data set, and the prediction accuracy is assessed using the validation set. The result are then averaged over all the splits. The advantage of this method over the K-fold cross validation is that the proportion of the training and validation set is not depends on the number of fold. The disadvantage of this method is that some observation may never be selected Into the validation set because there's randomization process. Whereas, some other samples may be selected more than ones into the validation set. In other words, validation set may overlap.



PREDICTIVE MODELING PIPELINE

1. Prediction Target

2. Cohort Construction

3. Feature Construction

4. Feature Selection

5. Predictive Model

6. Performance Evaluation

29. To conclude, in this lesson, we introduced the key steps in building a predictive model. Which include define what is the prediction target and construct the right patient cohort, then construct all the possible relevant features from data, then find which features are relevant, and view the predictive model, and finally, evaluate the model performance. Now you should be able to design a high level predictive modeling study using this pipeline on the HR data.