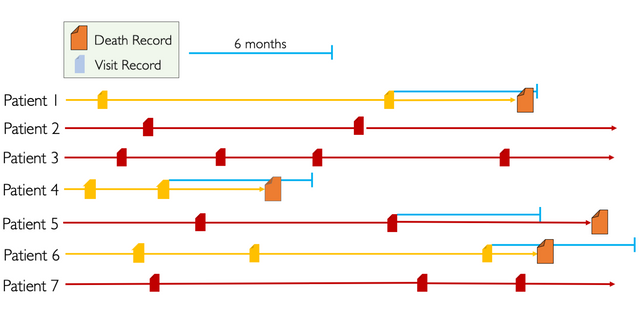
**Project 1: Patient Mortality Prediction**

**Task: Given all the past Electronic Health Records of a patient, predict whether a patient will pass away within 180 days following his/her last (available) exam.**

We define the True Positives as patients who pass away within 180 days of their last visit in the Data (Figure 1, Patients 1, 4, and 6). We define True Negatives as patients who have not passed away (Patients 2, 3, and 7) or who passed away more than 180 days after their last visit (Patient 5).

****

This project is based on the EHR Dream Challenge 2019 organized by the University of Washington and Sage Bionetworks: <https://www.synapse.org/#!Synapse:syn18405991/wiki/589657>

**Aim:** Develop an end-to-end machine learning pipeline to build and evaluate machine learning models on EHR data. You will use EHR data to predict mortality six months after the last visit.

Ideally, during this project you will learn how to:

* Explore an EHR dataset.
* Perform feature engineering i.e. extract features from the EHR for training and testing
* Join multiple datasets to create a coherent data frame for machine learning.
* Apply different machine learning algorithms.
* Evaluate and choose a machine learning model.

**Dataset:**

Since the Challenge Data cannot be made available for download without proper IRB approvals, we are making a synthetic dataset available. The Synthetic Public Use File (Synpuf) is a synthetic dataset created by the Centers for Medicare and Medicaid Services to provide a realistic example of claims data while protecting patient privacy. This data was converted to the OMOP Common Data Model by the CMS Working Group of the Observational Health Data Sciences and Informatics (OHDSI) community. We are making a portion of the full dataset available to participants for development and model testing purposes.

Dataset format: OMOP Common Data Model (<https://ohdsi.github.io/CommonDataModel/>)

Concept codes: <https://athena.ohdsi.org/search-terms/start>

Dataset can be downloaded from the Georgetown Box [here](https://georgetown.box.com/s/f4wx55mz9r8tooolynyozhc5is8emkve):

The dataset has been divided into two sets: training (for training and selecting the best Machine learning model) and testing (for evaluating for best model on a new unseen dataset). Each set contains several CSV files for different aspects of the EHR data (conditions, persons, visits, drugs, etc) in OMOP format.

**Features:**

Manipulate the dataset to extract features, which might be predictive of mortality. Some examples:

1. Demographics: use the person table to add demographics (Age, Gender, Race)
2. Conditions: life-threatening conditions and that might be predictive of very short survival (e.g.: renal failure, cancer, COPD, aids, etc.)
3. Compute Indexes: In medicine, the use of scoring systems to categorize patients into different risk strata is quite common. For example, the Charlson Comorbidity Index (CCI) ([Charlson et al., 1987](https://www.ncbi.nlm.nih.gov/pubmed/3558716)) quantifies an individual’s burden of disease and corresponding 1-year mortality risk. [CCI on MDCALC](https://www.mdcalc.com/charlson-comorbidity-index-cci)

**Train models, feature selection, and optimize hyperparameters:**

1. Use the sklearn package (<https://scikit-learn.org/stable/>) to train various classification models such as logistic regression, SVM, random forest.
2. Perform feature selection i.e. investigate if a smaller set of relevant features improve performance as opposed to using all features. See <https://scikit-learn.org/stable/modules/feature_selection.html> for more information
3. Use sklearn’s GridSearch (<https://scikit-learn.org/stable/modules/grid_search.html>) to optimize parameters for each model.

**Additional: Feature importance**

Use the random forest classifier (<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>) trained on the training data to extract the most important features for mortality prediction. Tutorial: <https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html>

**Evaluation**

Since this is a classification task (mortality with 6 months or not), report standard classification evaluation metrics (Precision, Recall, F1-score, AUC-ROC) on test set:   
  
**Additional resources and tutorials will be posted based on discussion with the instructor (Samir)**