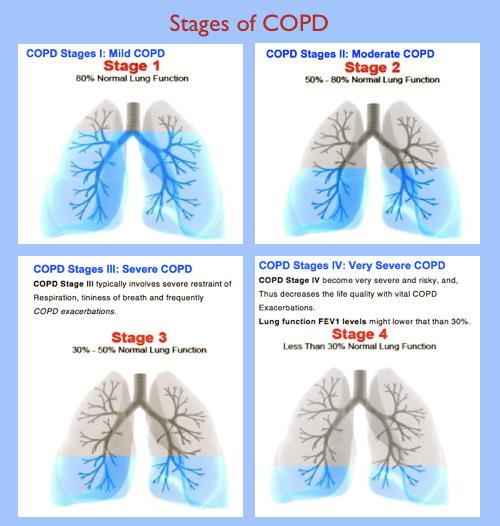
**Sample Project Plan**

The following plan contains specification around Chronic Obstructive Pulmonary Disease (aka COPD) as an example, however each candidate may pick different disease area and execute the entire case``````````````````

**Step 1: Problem statement & Hypothesis**

1. Download the required sample data (if data is already hosted by anyone in the batch and try to see if you can avail the same, rather than re-loading the same) and store it on GCP cloud storage/vm storage for further use.
2. Merge "Beneficiaries" , "Inpatient" and "Outpatient" data and run data profiling to understand available variables.
3. Choose co-morbidity/disease to analyze for re-admissions
4. Provide "Why","What" and "How" about the chosen disease.
5. Any statistical hypothesis can be tested for justify your chosen disease./Provide research paper for any references.
6. Understand the "DISEASE" in details for example -COPD -
   1. Do we know pre-conditions results in COPD i.e. respiratory infection, chronic bronchitis, Ashtma
   2. Do we know possible stages of disease in COPD

​

3. Do we understand possible proxy indicators for each of category of disease and how do we distinguish amongst these categories

7.  Utilize some of references as well to expedite your analysis. Jupyter notebook for data cleansing and  EDA part: <https://github.com/Jack-Etheredge/Predicting-early-hospital-readmissions/blob/master/01-Readmissions-Data-Cleaning-Pickling-ICD9-Codes-Living-JNE.ipynb>

8. Some of these readmissions example can provide good insights - <https://github.com/Jack-Etheredge/Predicting-early-hospital-readmissions/blob/master/Predicting%20patient%20readmission%20-%20JNE.pdf>

**Step 2: Data Engineering**

1. Finalize logic for index date and reasoning for the same. It should align to "Problem Statement" i.e. If readmissions for COPD to be predicted, we need to look at index hospitalizations and choose first index, second index etc..

​      2. How to choose index date is given for COPD patient.  Few examples

<https://www.sciencedirect.com/science/article/pii/S0954611113001868>

<https://journal.copdfoundation.org/jcopdf/id/1266/Validation-and-Assessment-of-the-COPD-Treatment-Ratio-as-a-Predictor-of-Severe-Exacerbations>

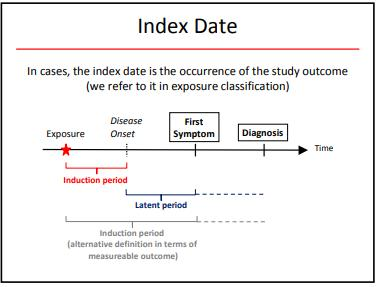
     3.  Few study design ideas for COPD

<https://journal.copdfoundation.org/jcopdf/id/1125/Identifying-Patients-With-COPD-at-High-Risk-of-Readmission> / <https://www.tandfonline.com/doi/abs/10.1080/15412555.2019.1688278?src=recsys&journalCode=icop20>

    4. Create patient cohort for further analysis by finalizing inclusion and exclusion criteria. Define all ICD codes/Procedure/CPT codes for disease categories. Also define particular category of disease to predict for readmissions.

   5.  Example of features for COPD are given at <https://www.nature.com/articles/s41598-019-39071-y>

   6. Index date journey example

​

**Example Criteria for COPD (Please define as per your use-case)**

Inclusion Criteria:

* Aged 40-75 years
* COPD group: baseline post-bronchodilator FEV1/FVC ＜ 0.7
* GOLD 0 group: individuals with chronic respiratory symptoms and/or high-risk factors
* Signed informed consent obtained prior to participations with the ability to comply with protocol and be available for study visits over 5 years

Exclusion Criteria:

* Acute exacerbation in the past 3 months
* Having other respiratory diseases with massive lung tissue destruction such as severe bronchiectasis and tuberculosis, etc
* The usual criteria of serious uncontrolled diseases
* thoracic or abdominal surgery in the last 3 months
* eye surgery in the last three months
* retinal detachment
* myocardial infarction in the last 3 months
* admission to hospital for any cardiac condition in the last month
* heart rate over 120 beats per minute
* antibacterial chemotherapy for tuberculosis
* pregnant or breast feeding

**Step 3: Feature Engineering**

* Define all features
* Calculate WOE (Weight of evidence) and IV (Information Value) for all features by binning them. This will be our feature selection methodologies. We will retain features with good IV >0.3 to 0.5.
* Verify which features are worth converting to categories than keeping at numerical. find out best transformation of features to provide best information gain in terms of outcome.
* Perform chi-square test on all features and verify their independence w.r.t. to outcome
* Run correlation only on selected features
* Finalize feature list, required transformation

**Step 4: Feature Engineering**

1. Decision on sampling - stratified sampling, random sampling, adaptive stratified sampling before data is split in train and test
2. Split data in train and test
3. Use maximum likelihood estimate (MLE) to find out coefficients of each feature (calculate manually rather than getting confused with any package)
4. If time permits, try to create decision tree by following the steps (however we will be revisiting decision tree in next project) <https://youtu.be/LDRbO9a6XPU>
5. "RiskSlim" Framework can be explored in terms of having to optimize risk scores.

**Step 5: Validation**

1. Apply on test data and calculate ROC-AUC
2. if ROC-AUC satisfactory, compare probability to observed frequencies by creating calibration plot.
3. If calibration plot comes nice, or refine probabilities.
4. Plot training data probability distribution density (aka pdf) plot and testing data probability density plot. Showcase mean and standard deviation in plot. Calculate similarity in distribution using t-statistics/K.S. statistics/overlapping index. if you see major drift in distribution, you may like to go back and refine your model and its probabilities.
5. Compare probabilities distribution by using K-L divergence/Bhattarchaya Distance. Can we use any other statistical distance to compare two distribution, please research and report. Be cautious on assumptions of using statistical distances.

**Step 6: Metrics and Cost Analysis**

1. Provide cost outcomes based on predictions and how providers can save cost by using predicted probabilities by your model.
2. Please merge prescription cost/drug cost to patient cohort.

**Step 7: Reporting and consolidation**

1. Create workflow to showcase cohort selection criteria (end to end)
2. Use "Model Automatic Documentation" packages available in python to document entire end to end modelling experiments.
3. Summarize cohort by providing distribution of data by age/gender/co-morbidity
4. Provide cost metrics
5. Deploy GCP based solution for real time predictions