Exploratory Data Analysis

Data Science: Drug Persistency

February 2023

LISUM16

Presented by: Olivia Foster, Sammy Suliman, and Tahsin Azad

Data Glacier

Your Deep Learning Partner

Problem Statement

One of the challenges faced by Pharmaceutical companies is the persistence of a drug (that is, the extent to which a patient will act in accordance with the prescribed time interval, and dose of a medication) as the physician prescribed it. To solve this problem ABC pharma company approached an analytics company to automate this process of identification. In this problem, we will automate the process of classifying factors that determine the persistence of a drug through Machine Learning and Python.

Drug persistence is a task of classifying different disorders and a patient's medical history to determine the dose and length of dose. In order to train our model, we will need to classify risk factors, medical histories, and disorders. To do this, we will be using a dataset based on over 3000 patients' records.

Data Analysis Approach

- Explore and understand the data.
- Prepare and clean the data.
- Analyze the data and find the features/variables that affects drug persistency.
- Give recommendations for the classification model that is to be built to automate the process of drug persistency identification.

Data Glacier

Your Deep Learning Partner

Data Exploration

- One file used for the dataset
- 3,424 data points
- 69 variables initially
  - 8 variables were dropped
  - 4 variables were derived/transformed
  - 61 variables were used for final analysis

Data Glacier

Your Deep Learning Partner

Overview of Cleaning Process

The following was done:

- Nulls and ‘unknown’ values were processed/eliminated
- Converted columns with Yes/No inputs into 1 or 0 inputs for ease of calculations.
- Non-numeric features were either dropped (ex. Patient_ID) or were hotkey encoded to have some numerical value for analysis.

Data Glacier

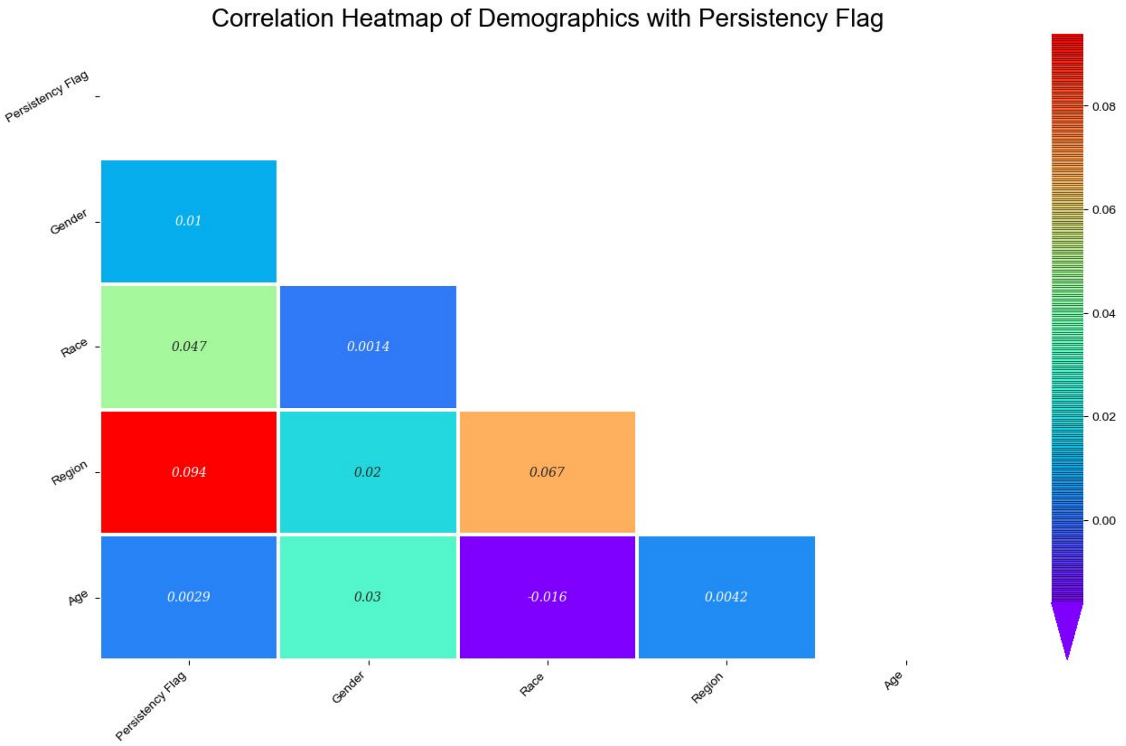
Your Deep Learning Partner

Correlation Analysis

Data Glacier

Your Deep Learning Partner

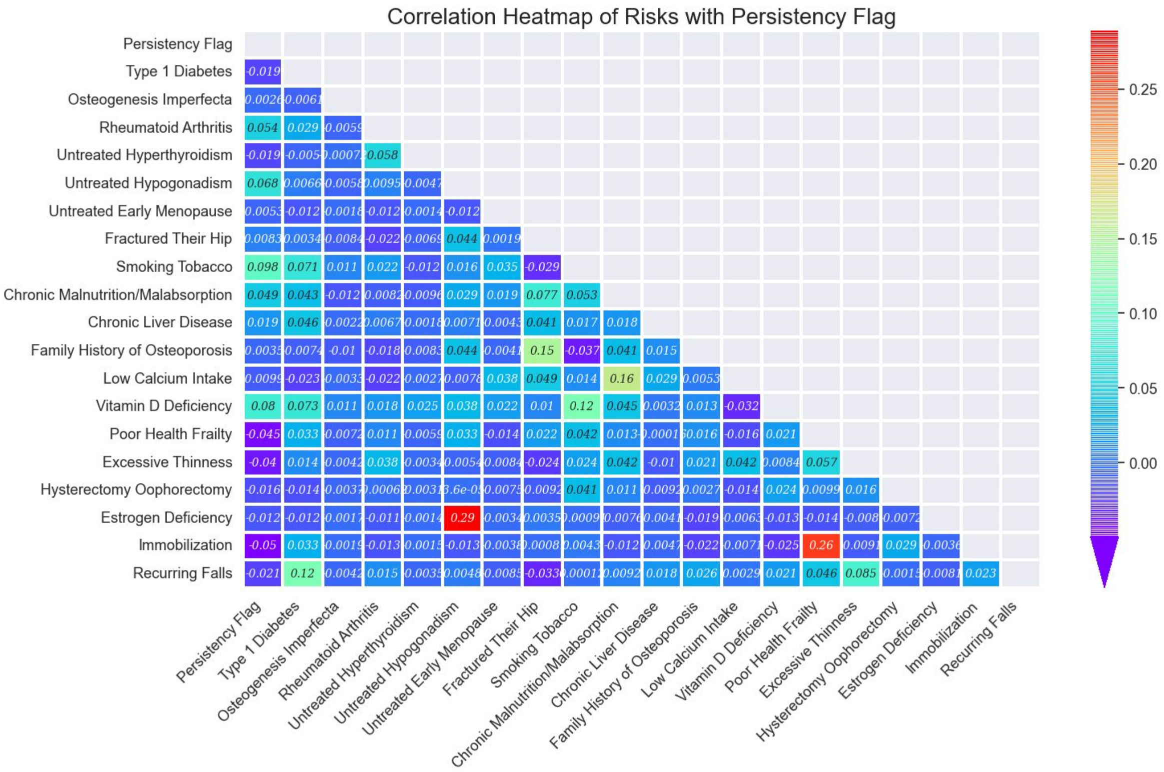
Heatmaps



Part of our analysis was determining what factors would have the greatest influence on each other, but more importantly on persistency.

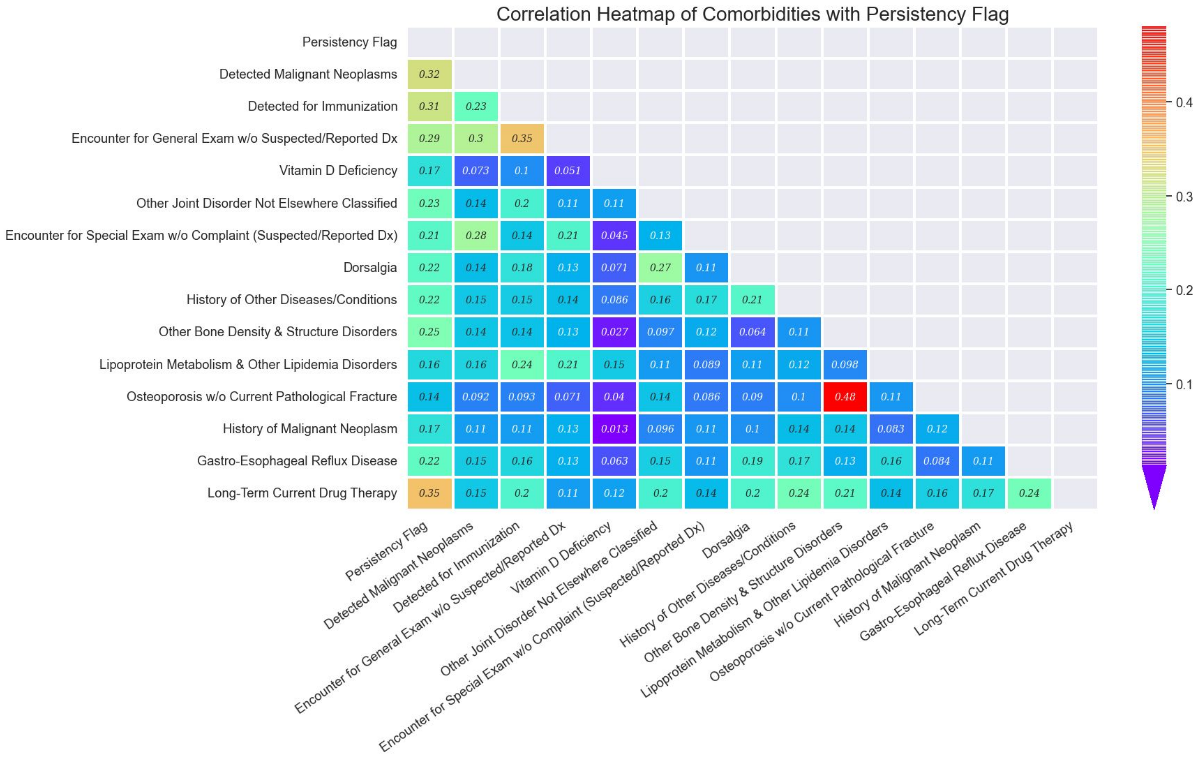
Our heatmaps should that influence. The closer to 1 the number is, the greater the influence the features have on each other.

Heatmaps



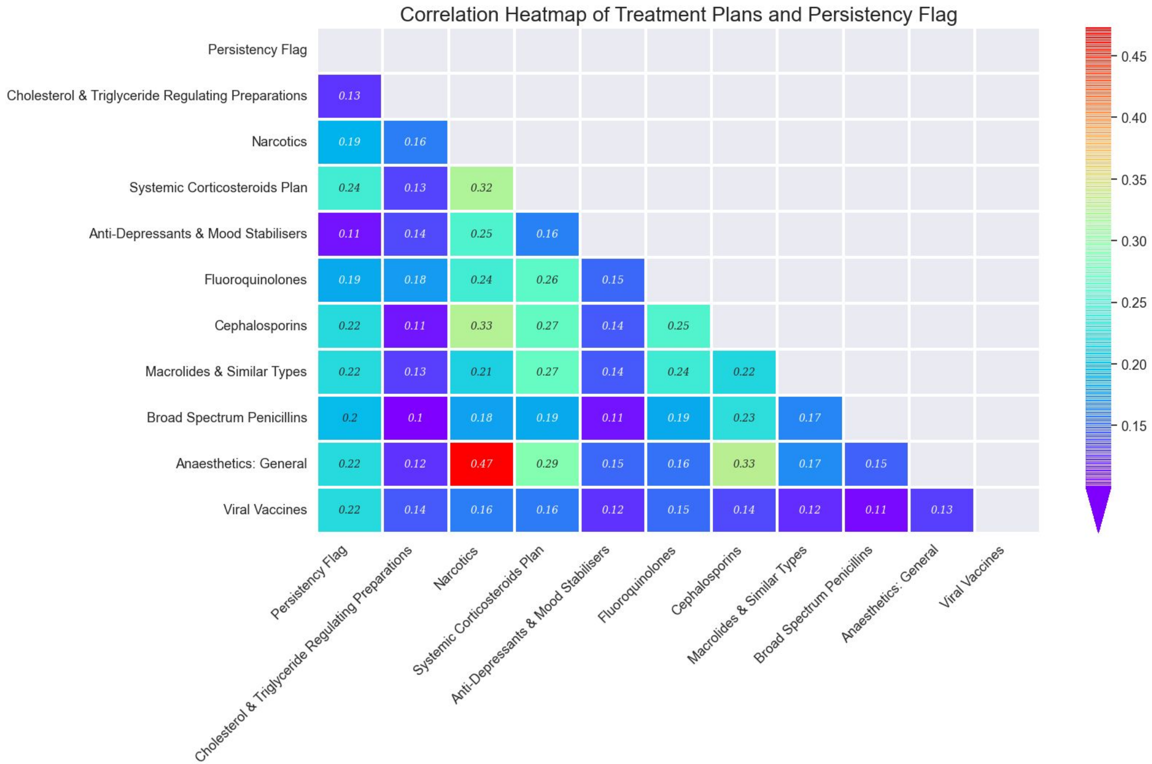
For risks, we see several points of strong correlation. But we're focusing most on Persistency. From our Risks, our strongest influencing factor on persistency is whether or not a patient smoked tobacco.

Heatmaps



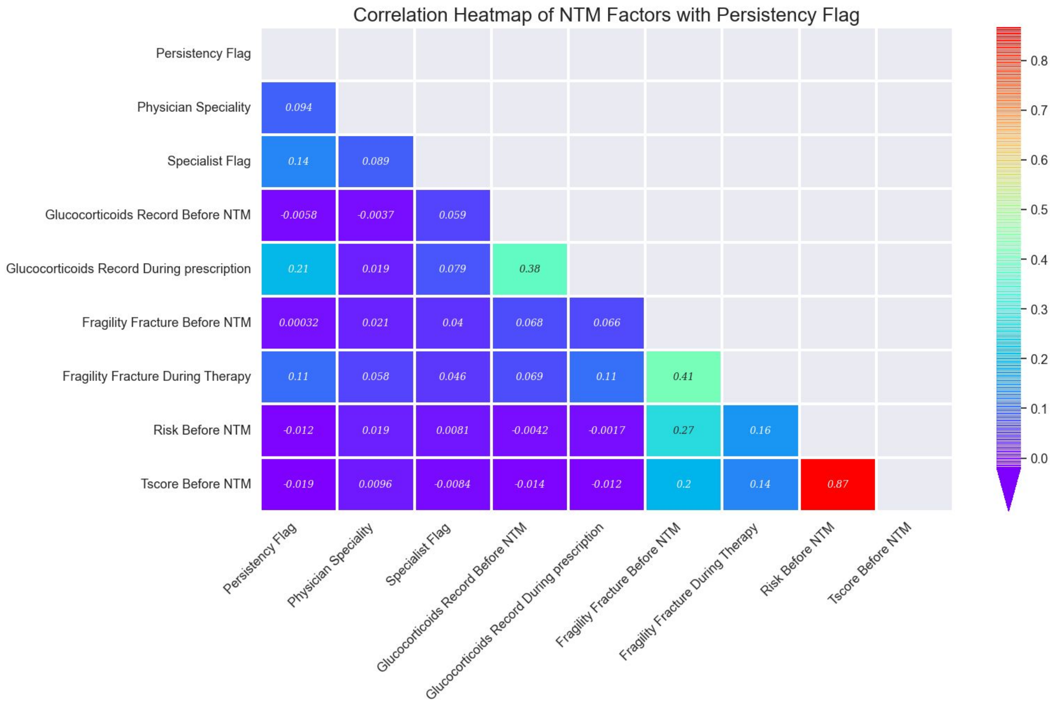
Similar to the previous slide, we see several factors having strong influences on one another in Comorbidities. Comorbidities have the greatest influence on persistency as we have 4 features (Detected Malignant Neoplasms, Immunization, General Exam without Suspected/Reported Diagnosis, and Long-Term Current Drug Therapy) with a correlation factor that is greater than 0.25.

Heatmaps



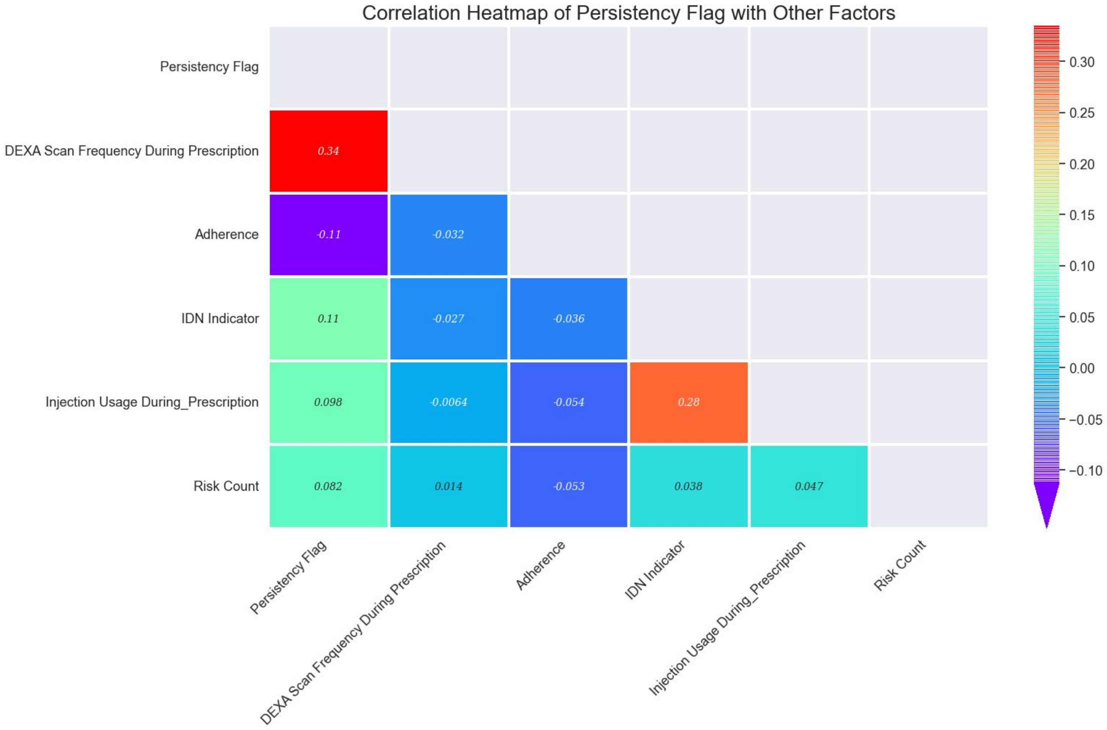
For treatment plans the greatest influence on persistency is the use of systemic corticosteroids plans. However, it should be noted that overall, treatment plans have a consistently higher influence on persistency than other categories.

Heatmaps



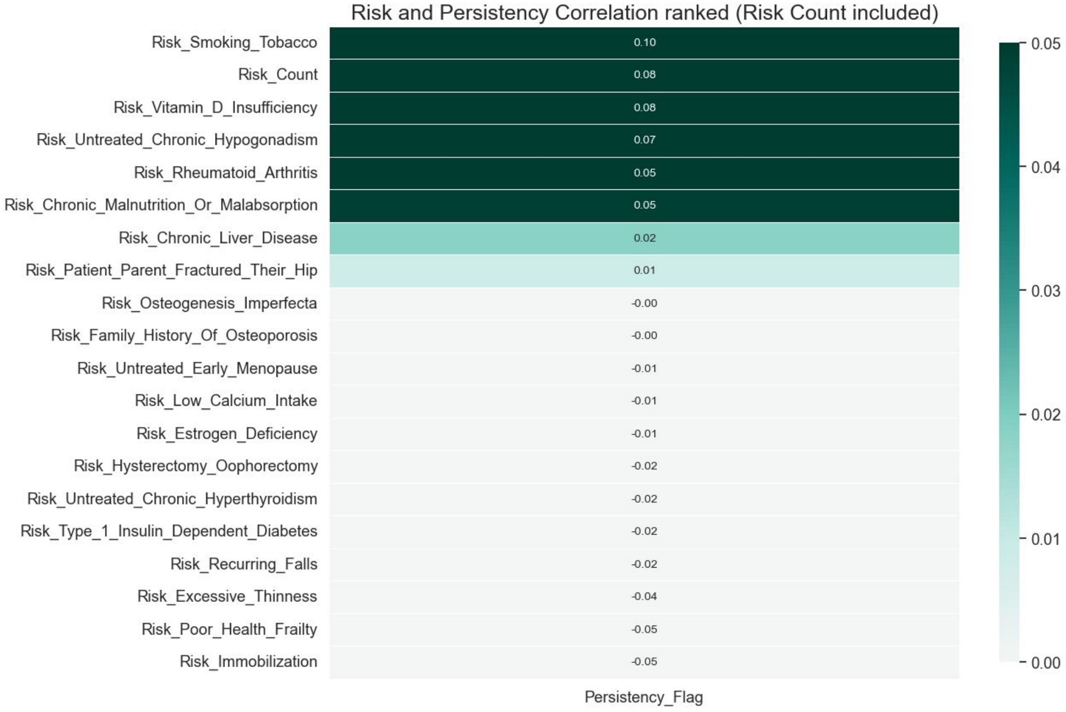
For NTM factors, overall they do not seem to have any strong influence on persistency except for glucocorticoids records during treatment, indicating that the drug is the most important part of the treatment.

Heatmaps



For factors that didn't really fit in any other category, there weren't any strong influences, barring DEXA Scan Frequency During Prescription, suggesting that regular check-ins play a huge part in determining how well a patient sticks with their regiment.

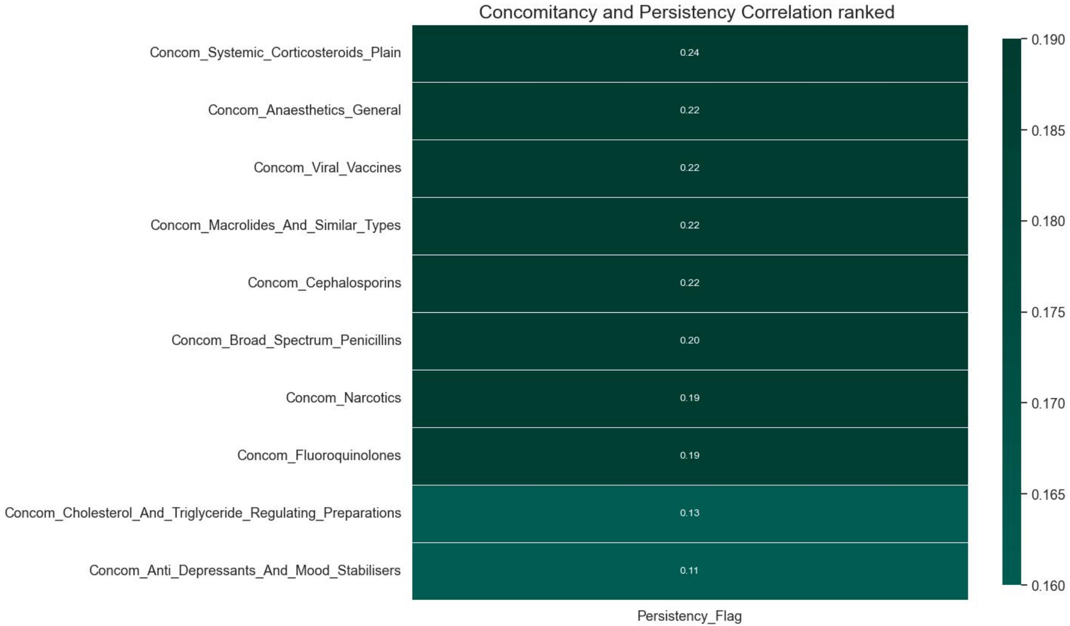
Correlation Ranking Compared to Persistency



To confirm our suspicions found in the heatmap, we constructed similar but different models in correlation rankings. This model ranks the correlation from highest to lowest.

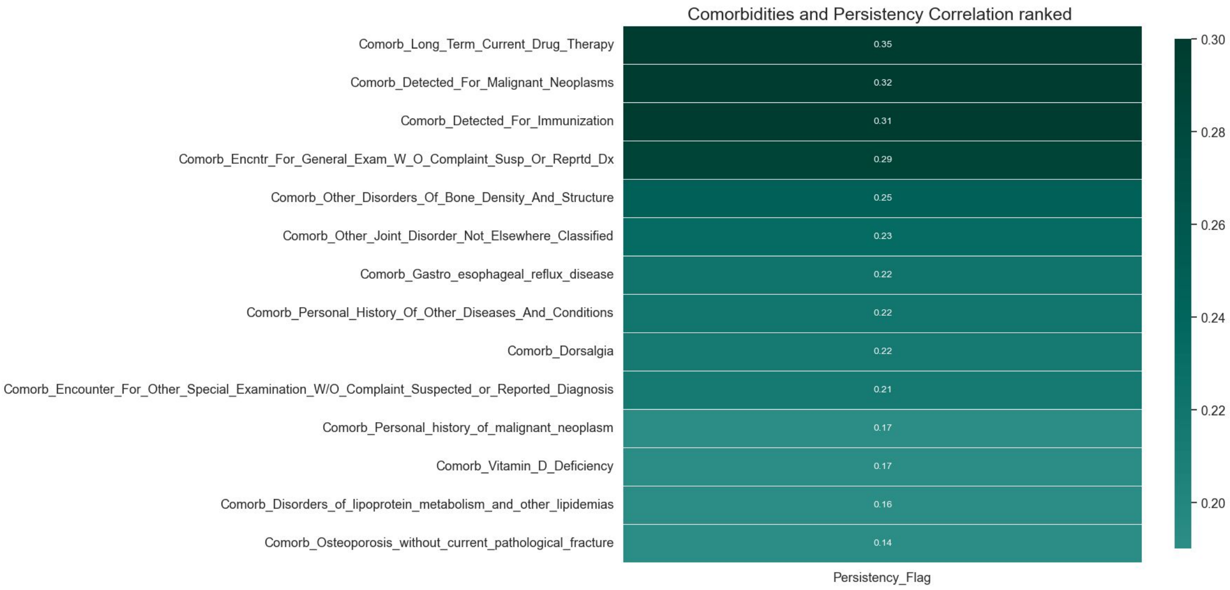
In our risks, we do see see that risks do not have a strong influence on persistency, with Smoking Tobacco being the highest at 0.10, which is the rounded number (actual number is 0.098).

Correlation Ranking Compared to Persistency



In treatments (also known as concomitancy) we find that the correlation rankings confirm our suspicions regarding treatments with systemic corticosteroids being the being most influential feature.

Correlation Ranking Compared to Persistency



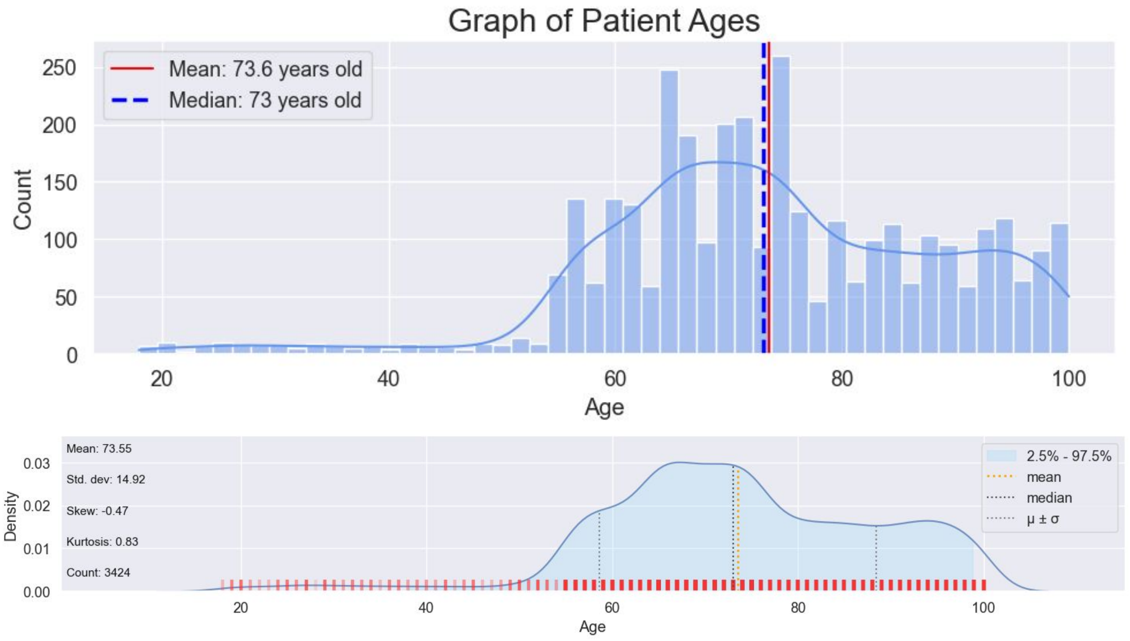
Here we can see that once again we find that Comorbidities have a very strong influence on persistency with the majority of the points being over 0.20.

Analysis of Features

Data Glacier

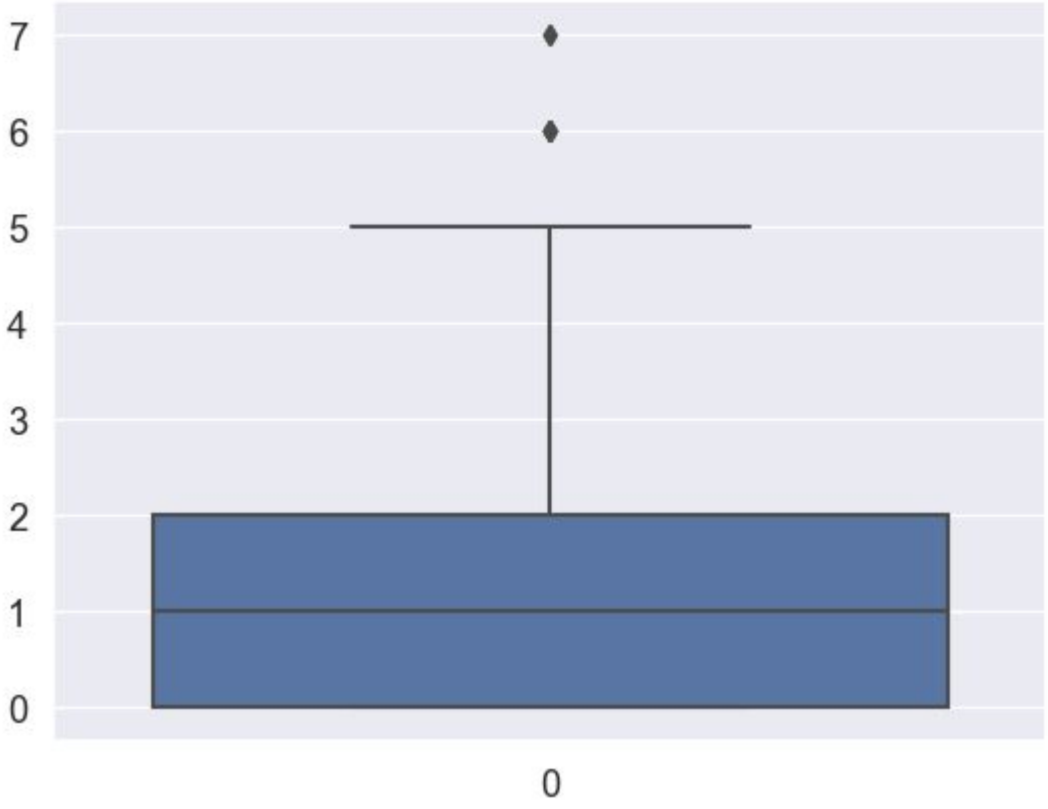
Your Deep Learning Partner

Patients' Ages



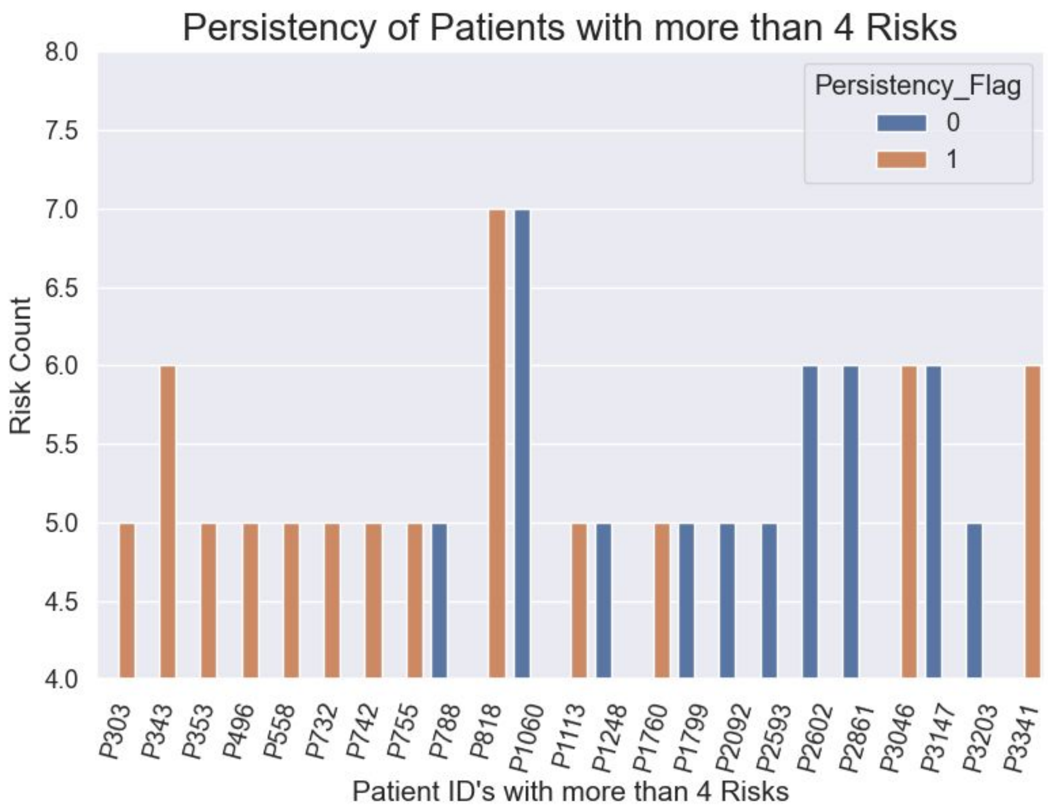
In our data, extrapolated ages based on the age buckets that were found in our raw data. Using two different models, we verified that our data produced consistently and applicable results. In this we founds that the average age of patients was almost 74 years old and the median age was 73, suggesting that there are no outliers heavily influencing our data.

Boxplot- Outliers for Number of Risks



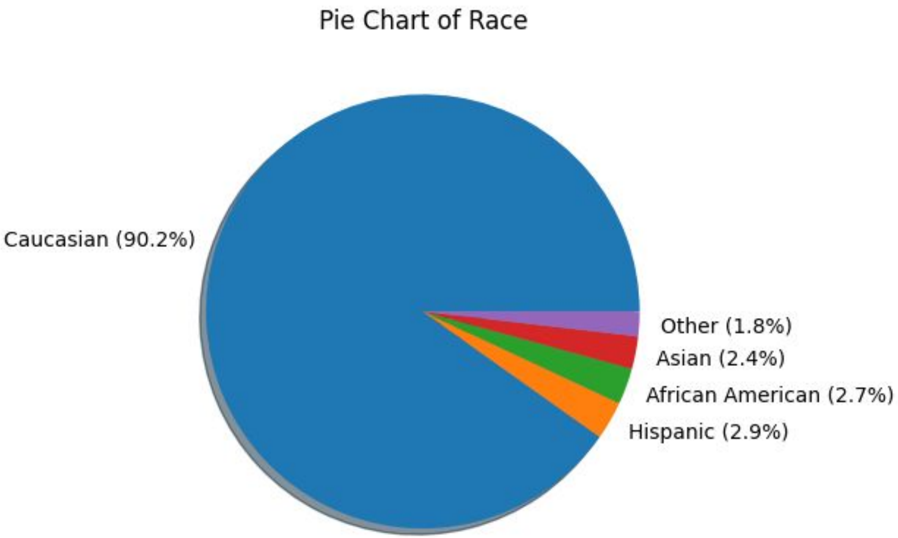
One important observation in analyzing the data is searching for outliers which may affect the data. Since most of the data is binary or categorical, only outliers for Number of Risks were observed. The boxplot represents this, showing that those with 6 and 7 risks were outliers compared to the rest.

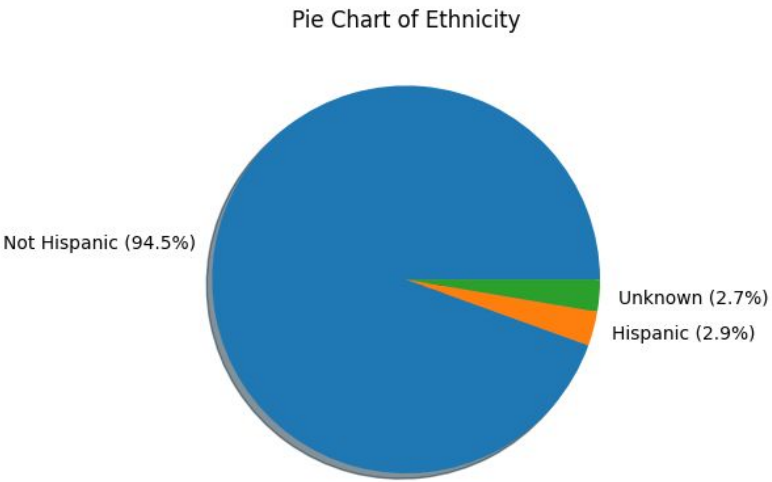
Persistency of Patients with more than 4 Risks



@Sammy: Could you do this description since you developed this code? -Livia

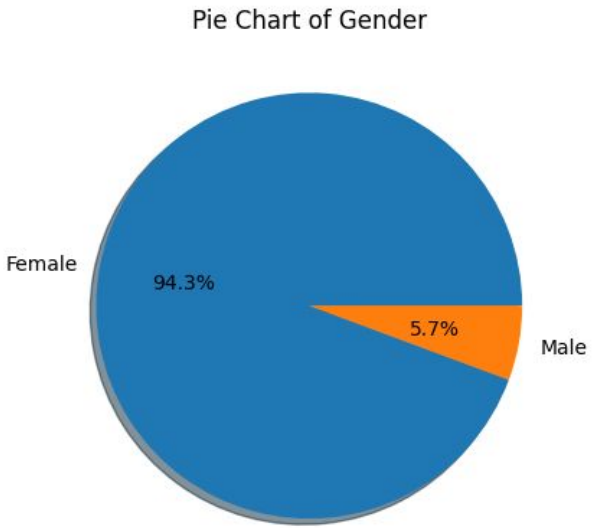
Demographics – Race and Ethnicity

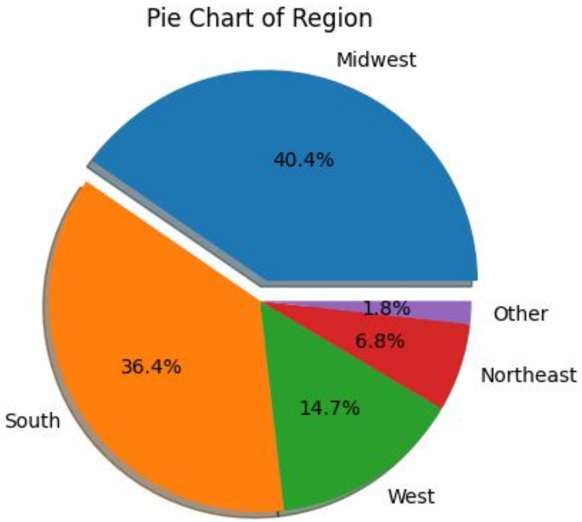




In terms of demographics, we can see that the majority of our data was collected from patients who were Caucasian and were not Hispanic. This could influence risks and comorbidities, which in turn could also influence the persistency.

Demographics – Region and Gender

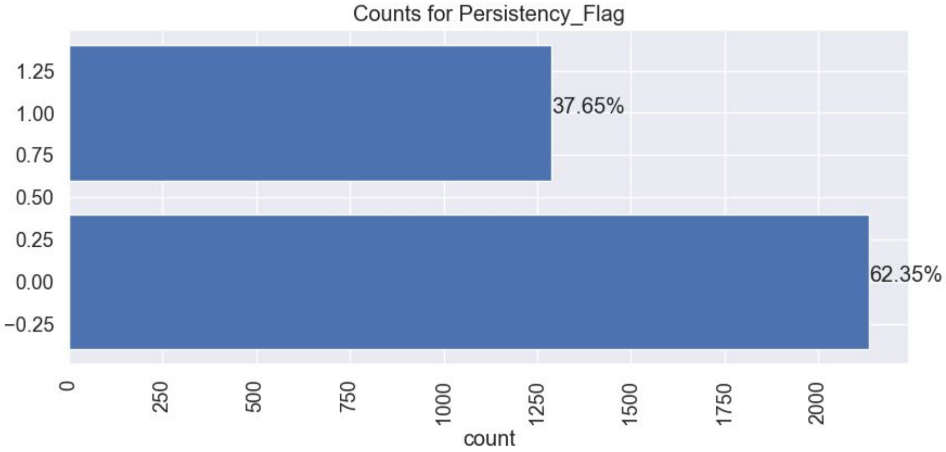


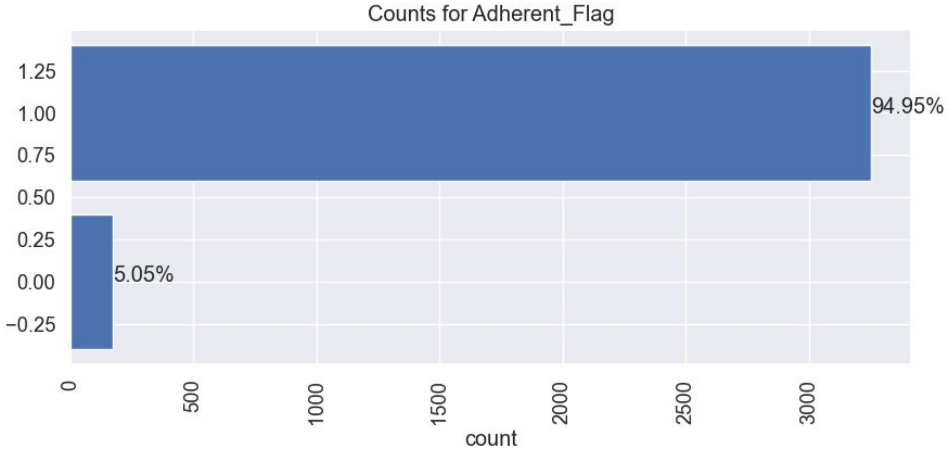


For region the majority of patients come from the South and Midwest (about 77%).

For the gender of patients, over 90% of our patients are female. Because we have a several features that are only relevant to female patients, this information is incredibly helpful and will help build an accurate model.

Demographics – Persistency and Adherence

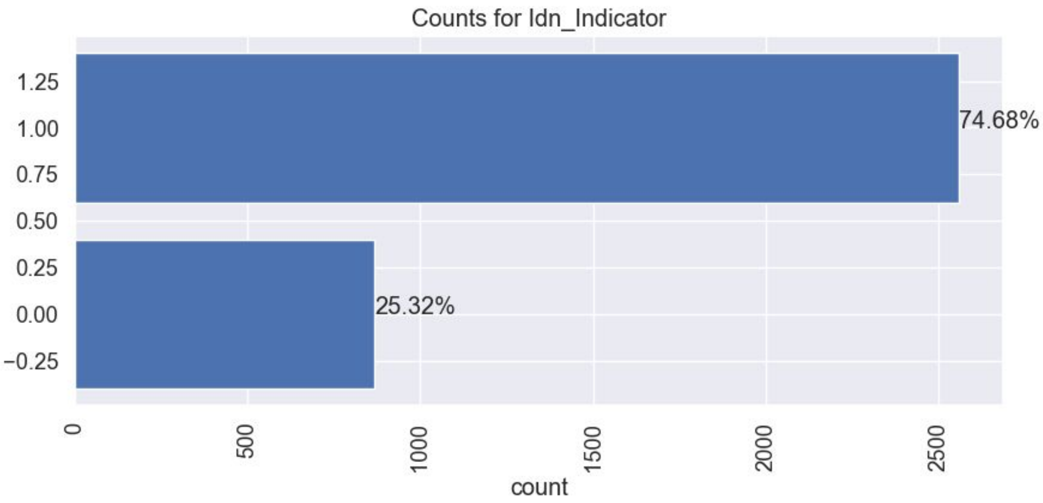




For persistence, our target, we find that the majority of patients are not persistent (0). However, this seems to correlate with the fact that many patients are negative to many features.

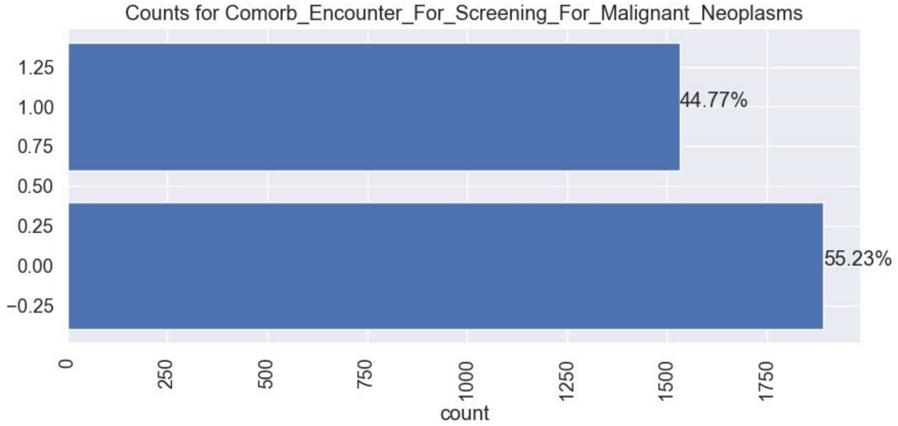
Interestingly though, patients seem to adhere to their medication regimen (1) at an almost 95% success rate. This suggests that if a treatment plan is relevant to a patient given their condition, they will stick to it.

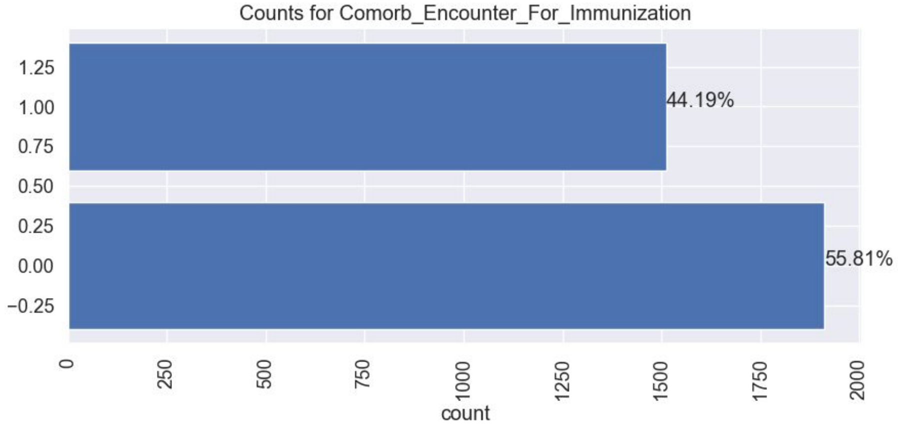
Demographics – IDN Indicator



The IDN indicator suggestions that the majority of patients' information was uploaded to the IDN system that allows doctors to share their records regarding their patients.

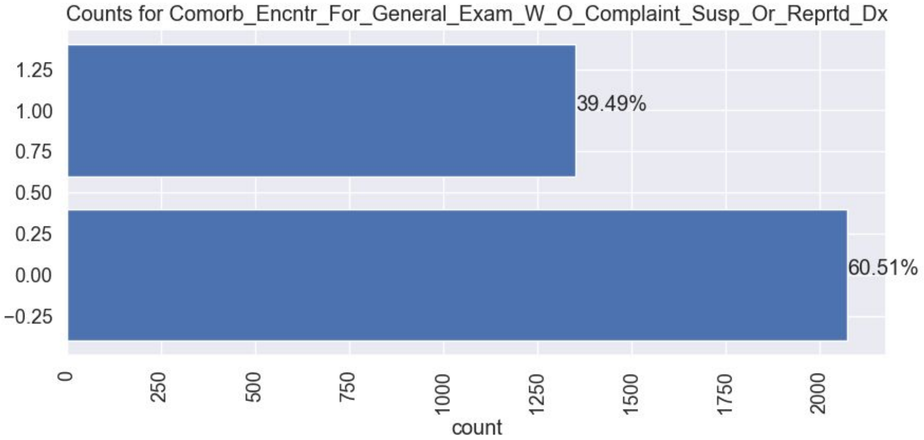
Comorbidities

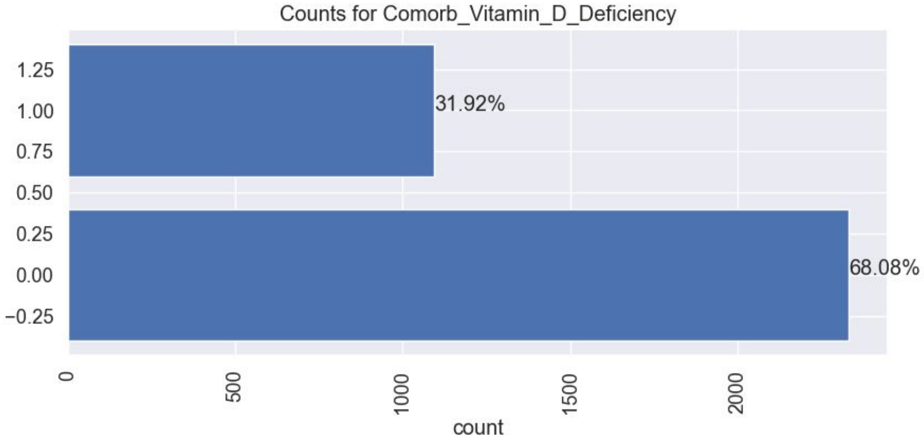




For both ‘comorbidities encountered for screening for malignant neoplasms’ and ‘comorbidities encountered for immunization’, a little more than half of the patients came back negative for both conditions.

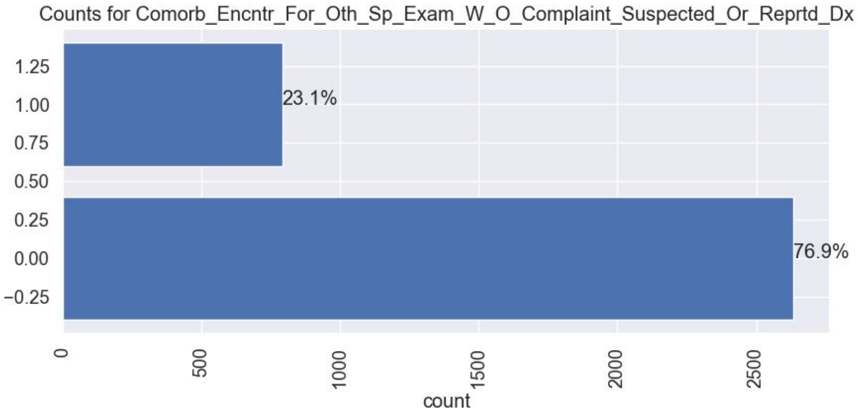
Comorbidities

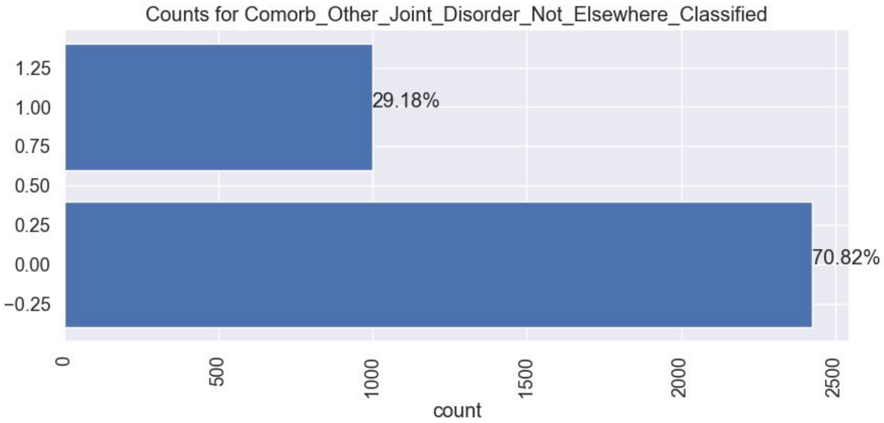




For both 'comorbidities encountered during general exam without complaint: suspected or reported during diagnosis' and 'comorbidities for vitamin D deficiency', between 60 and 68% of the patients came back negative for both conditions.

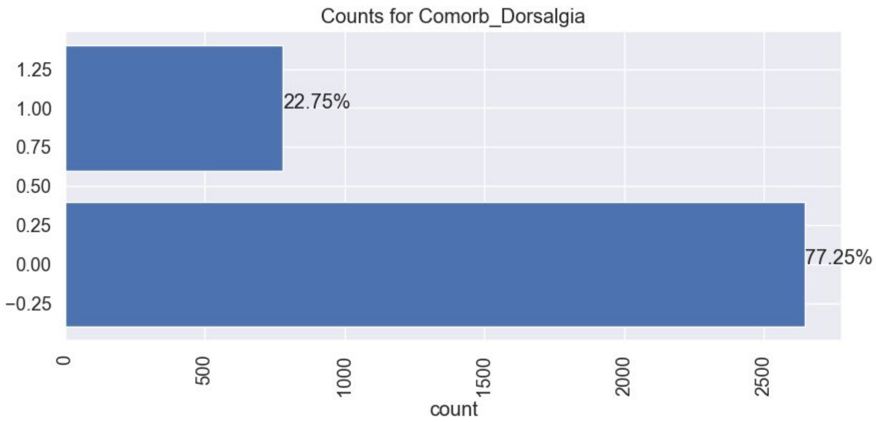
Comorbidities

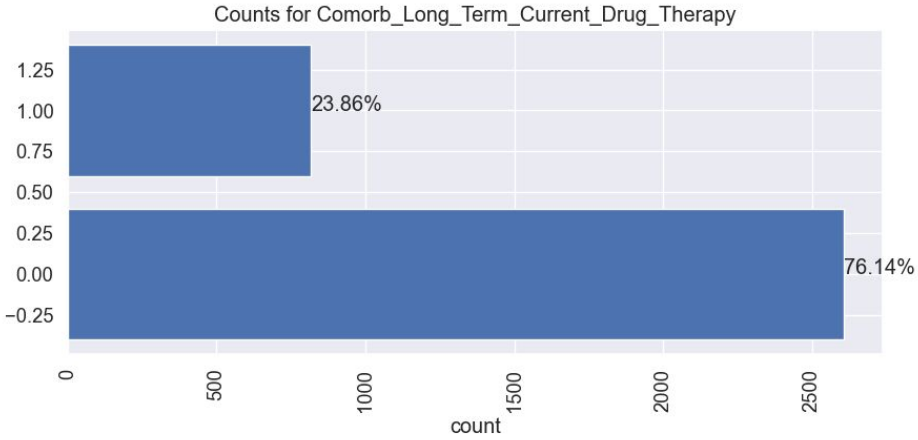




For both ‘comorbidities encountered during special exam without complaint: suspected or reported during diagnosis’ and ‘comorbidities for other joint disorders not elsewhere classified’, between 70 and 77% of the patients came back negative for both conditions.

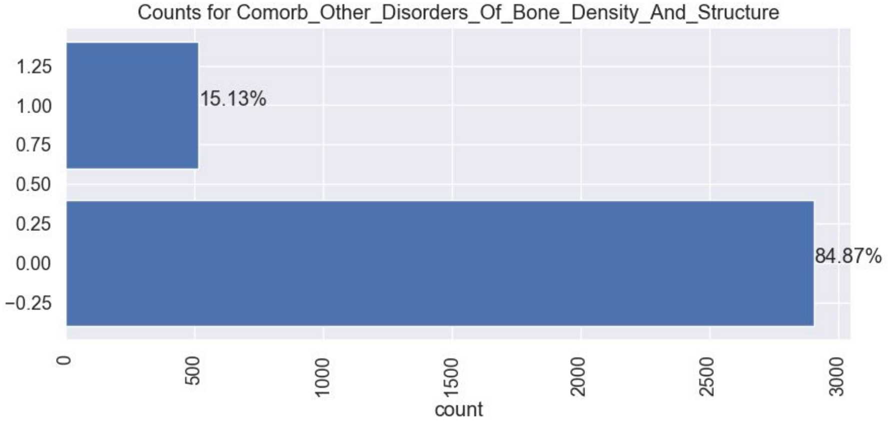
Comorbidities

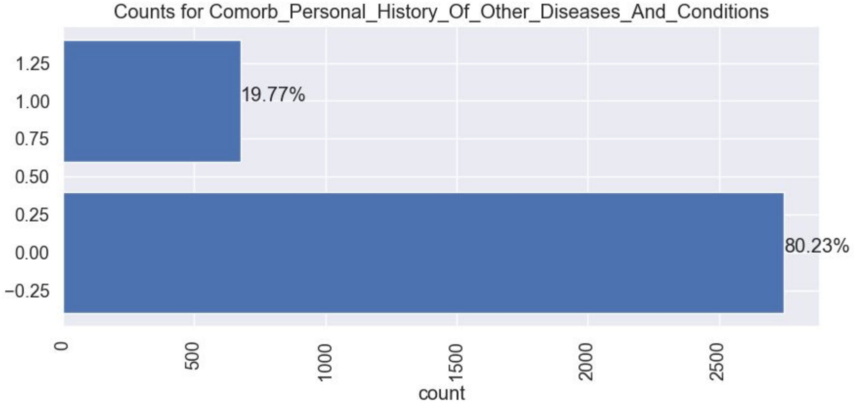




For both ‘comorbidities for dorsalgia’ and ‘comorbidities for long-term current drug therapy’, between 76 and 78% of the patients came back negative for both conditions.

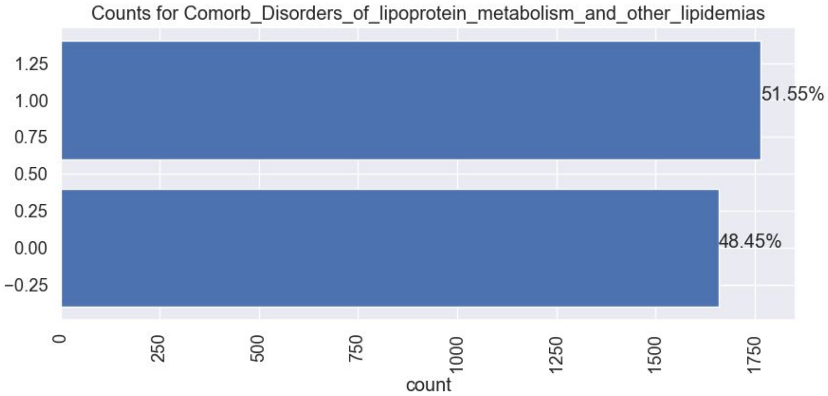
Comorbidities

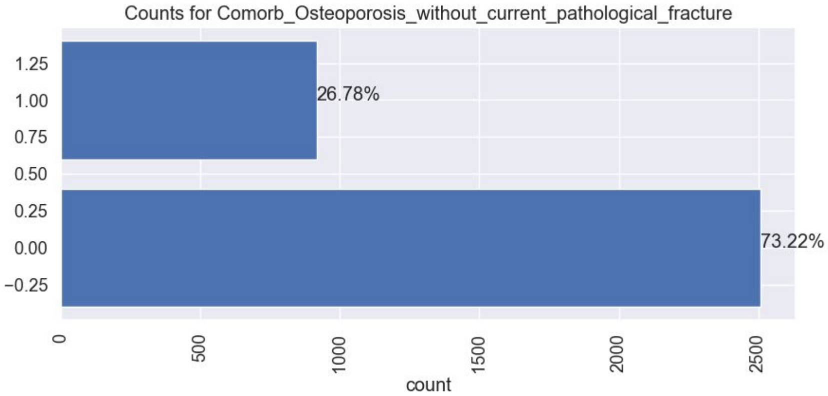




For both ‘comorbidities for other disorders of bone density and structure’ and ‘comorbidities for personal history of other diseases and conditions’, between 80 and 85% of the patients came back negative for both conditions

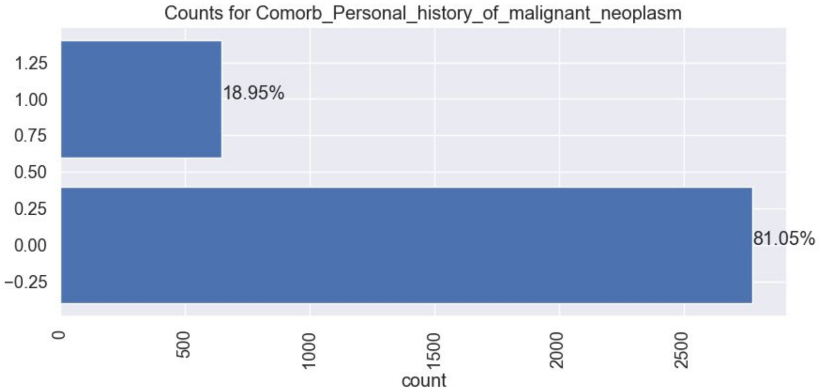
Comorbidities

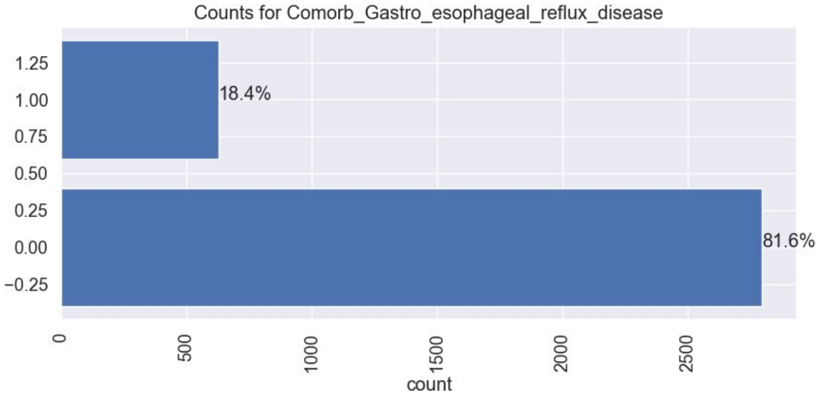




Here, we see something a little different in that for 'comorbidities for disorders of lipoprotein metabolisms and other lipidemias' we actually have a slightly higher positive outcome than negative. For 'comorbidities for osteoporosis without a current pathological fracture' however, roughly 73% of patients came back negative.

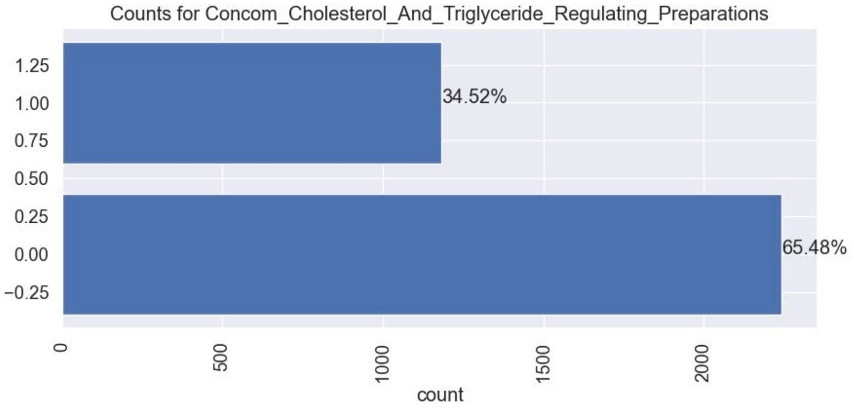
Comorbidities

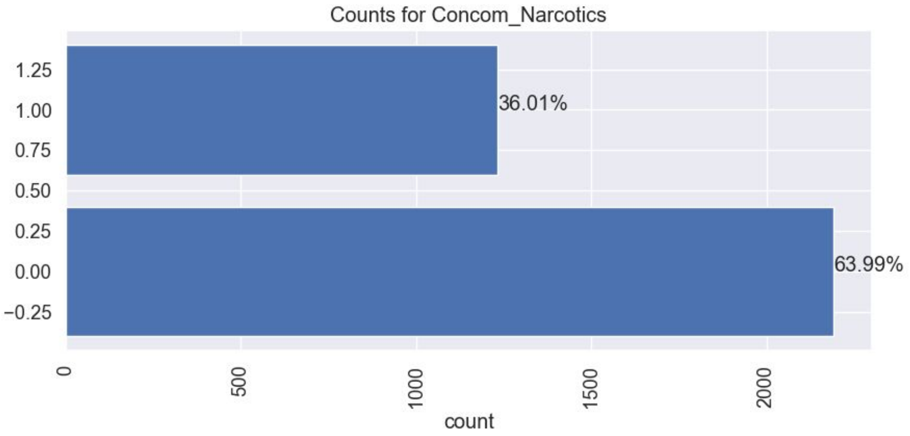




For both ‘comorbidities for personal history of malignant neoplasm’ and ‘comorbidities for gastroesophageal reflux disease’, roughly 81% of the patients came back negative for both conditions.

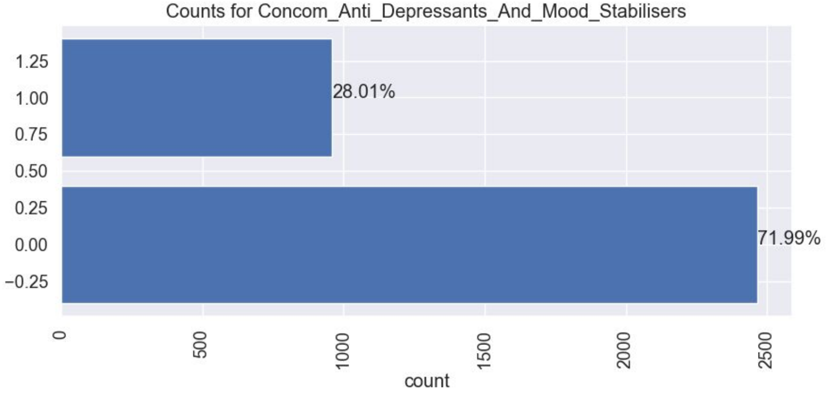
Concomitancy

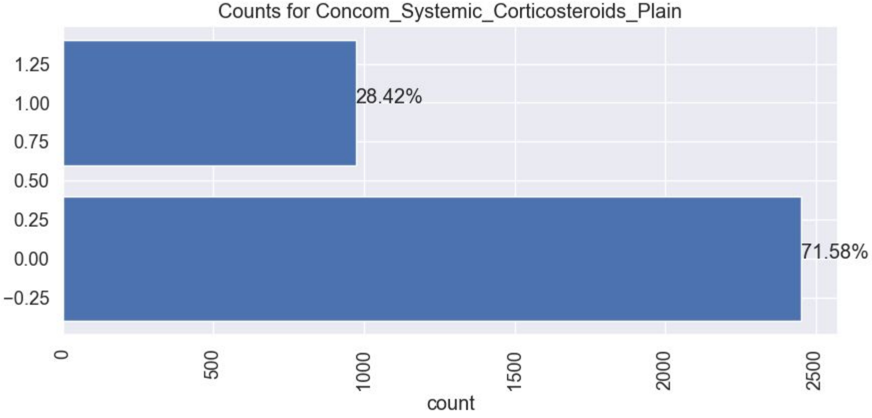




For both ‘concom_cholesterol_and_triglyceride_regulating_preperations’ and ‘concom_cephlaospirins’, majority of the patients came back negative for both conditions, both being between 63 and 66%.

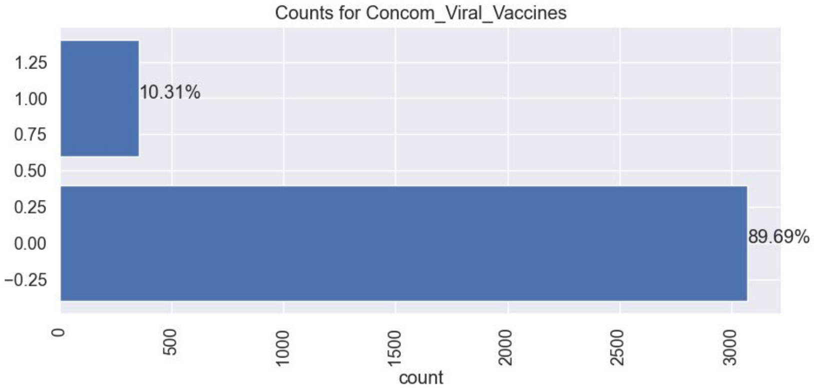
Concomitancy

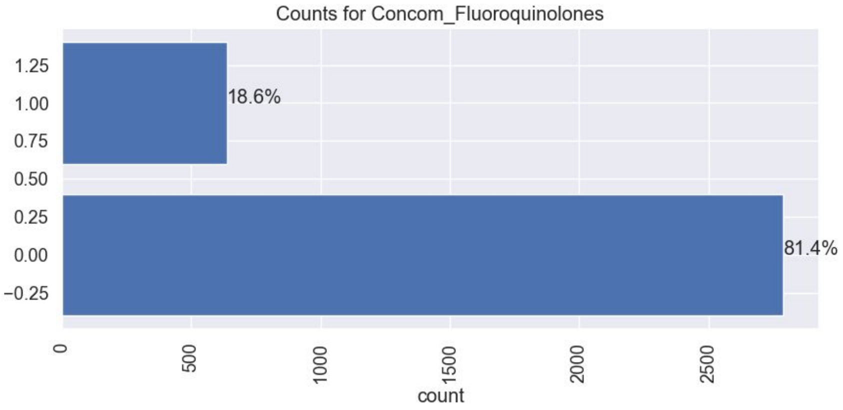




For both 'concom_anti_depressants_and_mood_stabilisers' and 'concom_systemic_corticosteroids', majority of the patients came back negative for both conditions, both being between 71 and 72%.

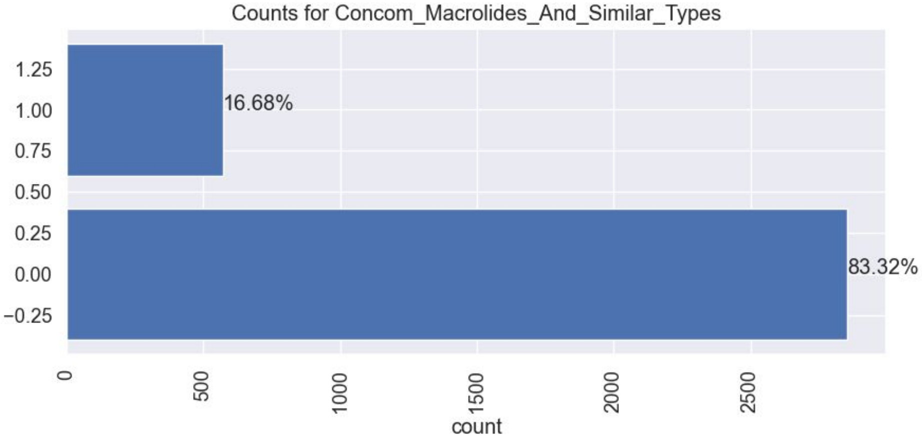
Concomitancy

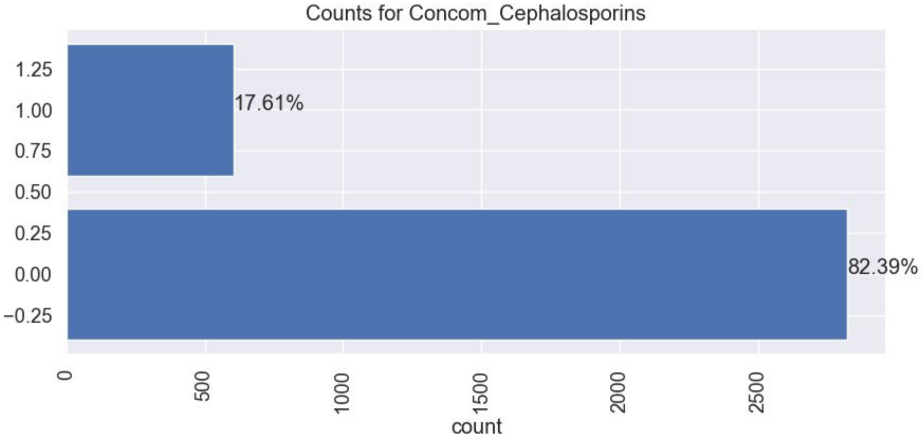




For ‘concom_viral_vaccines’, an overwhelming majority of patients came back as negative. For ‘concom_fluoroquinolones’, majority of the patients came back as negative, with 81.4%.

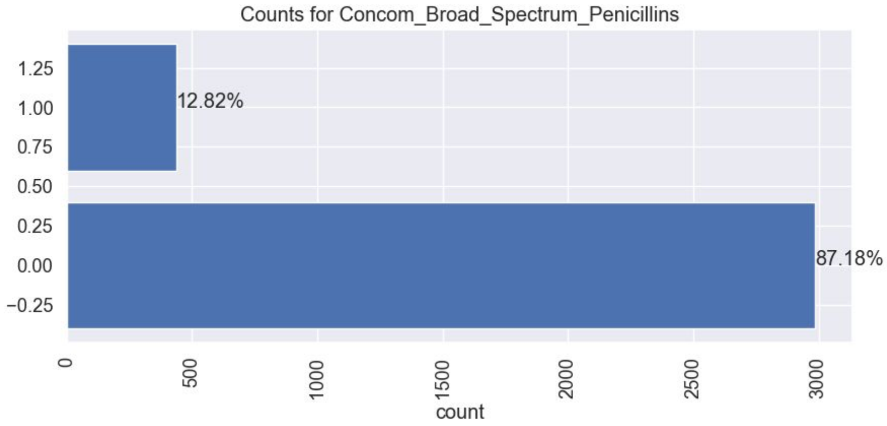
Concomitancy

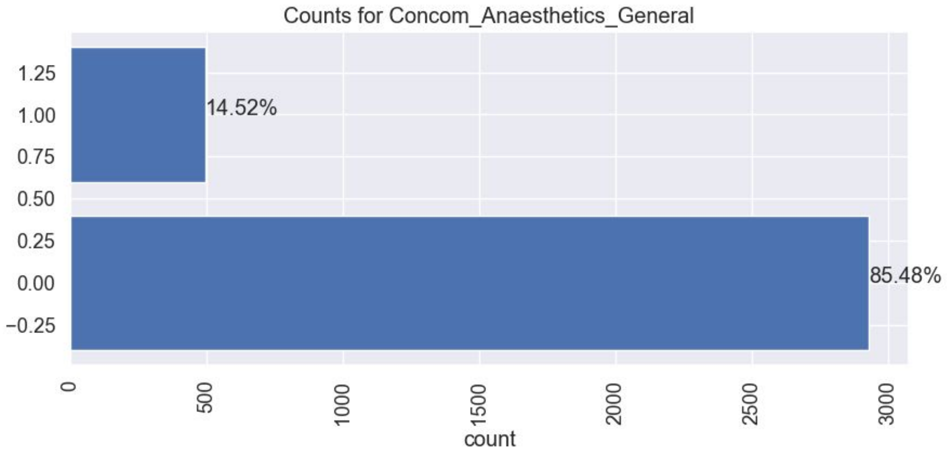




For both ‘concom_macrolides_and_similar_types’ and ‘concom_cephlaospirins’, majority of the patients came back negative for both conditions, both being between 82 and 84%.

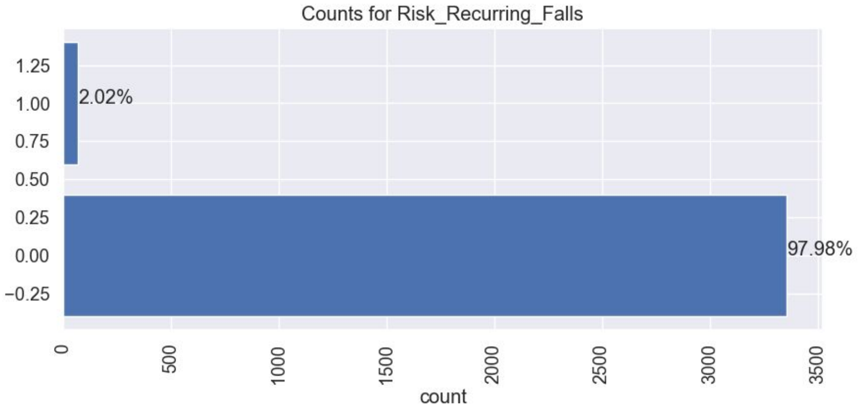
Concomitancy

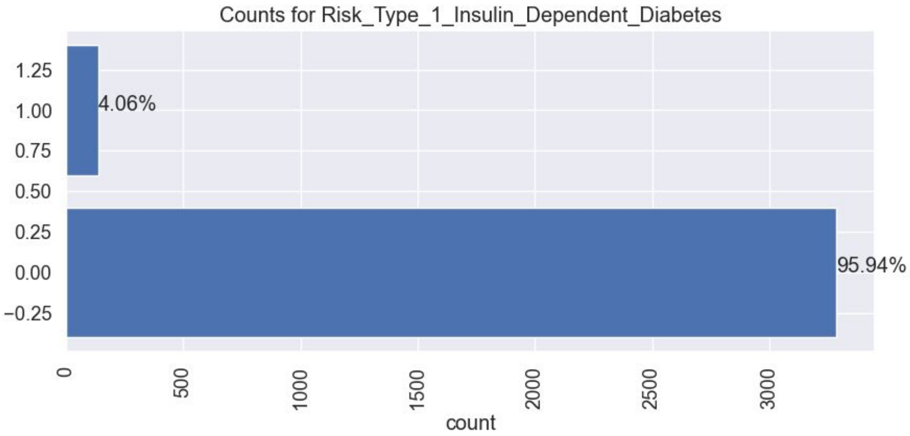




For both ‘concom_broad_spectrum_penicillins’ and ‘concom_anaesthetics_general’, an overwhelming majority of the patients came back negative for both conditions.

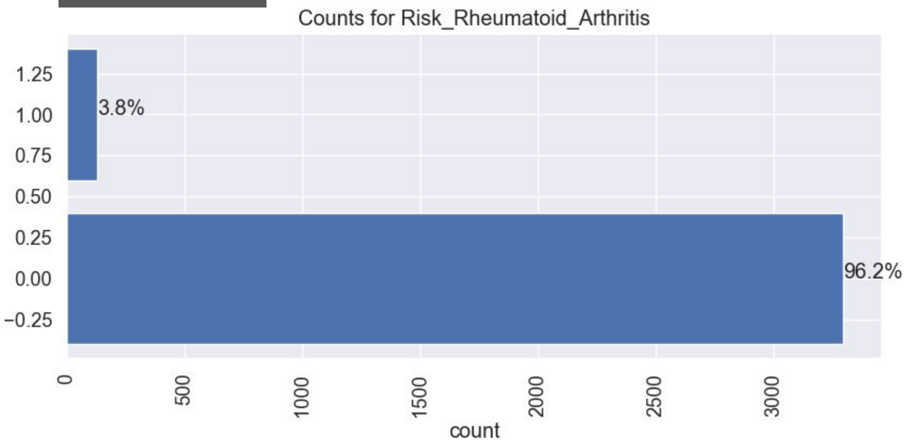
Risks

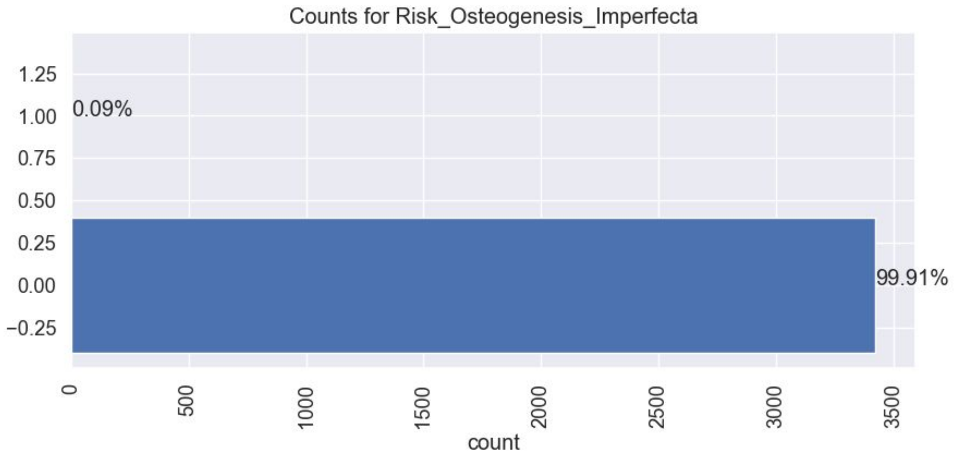




For both ‘risk_recurring_falls’ and ‘risk_type_1_insulin_dependent_diabetes’, nearly all of the patients came back negative for both conditions.

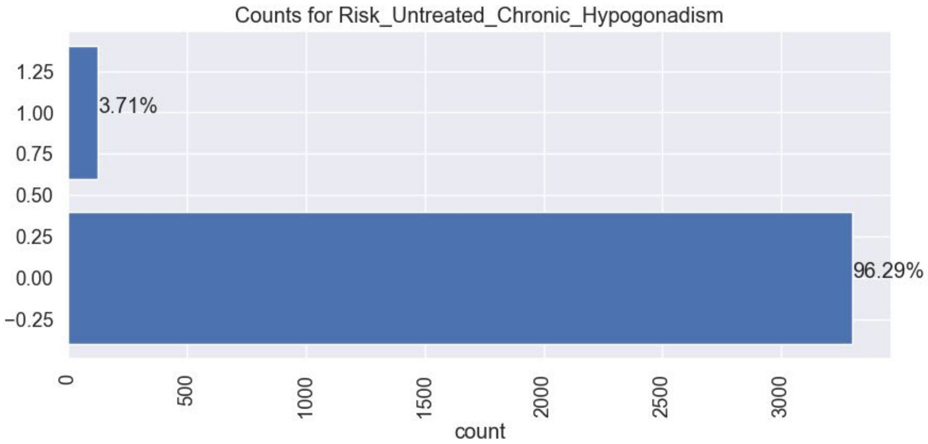
Risks

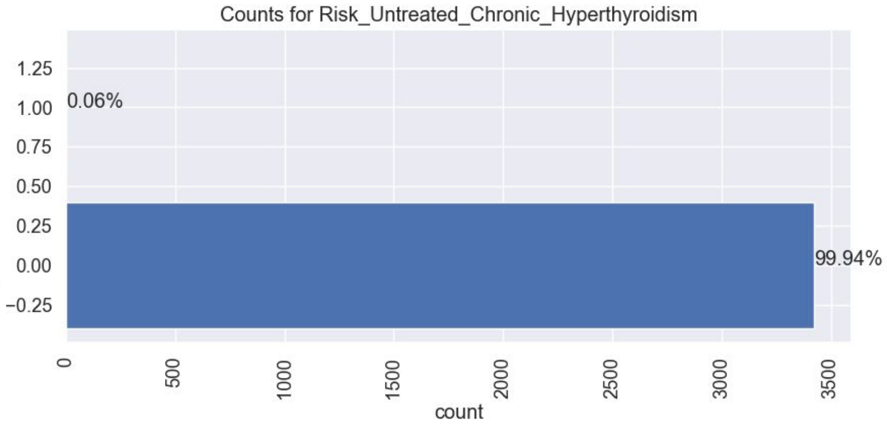




For ‘risk_rheumatoid_arthritis’, an overwhelming majority of the patients, 96.2%, came back as negative. For ‘risk_osteogenesis_imperfecta’, nearly all of the patients came back at negative with only 0.09% resulting in positive.

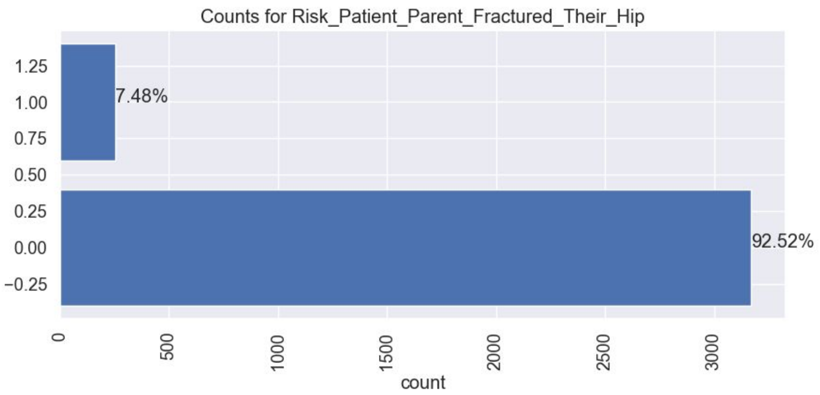
Risks

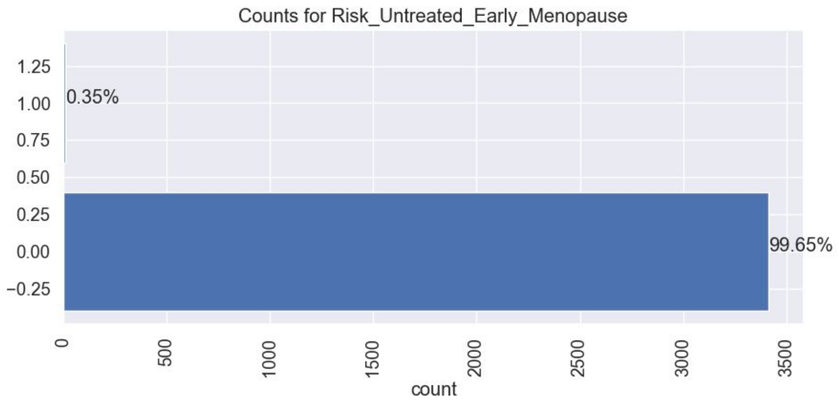




For ‘risk_untreated_chronic_hypogonadism’, an overwhelming majority of the patients, 96.29%, came back as negative. For ‘risk_untreated_chromic_hyperthyroidism’, nearly all of the patients came back at negative with only 0.06% resulting in positive.

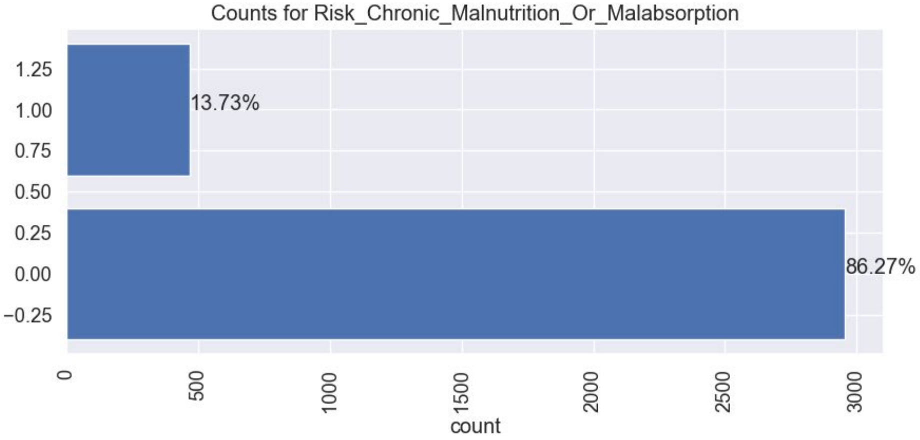
Risks

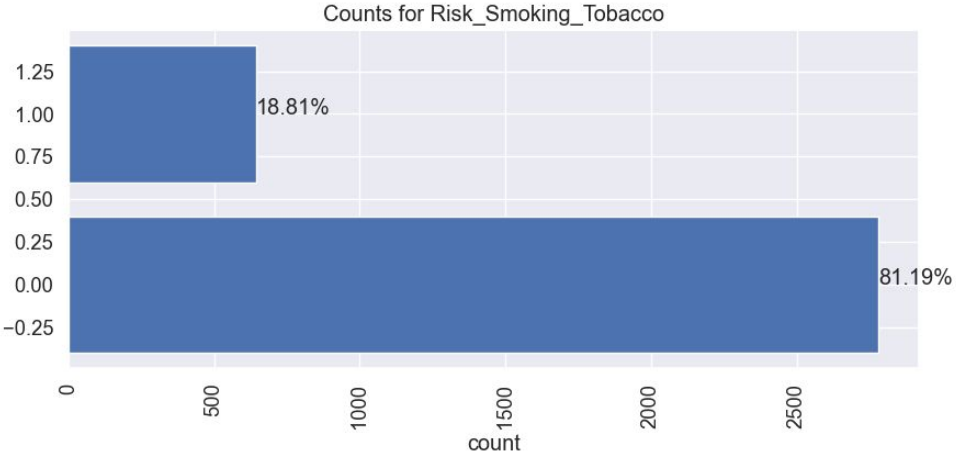




For 'risk_patient_parent_fractured_their_hip', an overwhelming majority of the patients, 92.52%, came back as negative. For 'risk_untreated_early_menopause', nearly all of the patients came back at negative with only 0.35% resulting in positive.

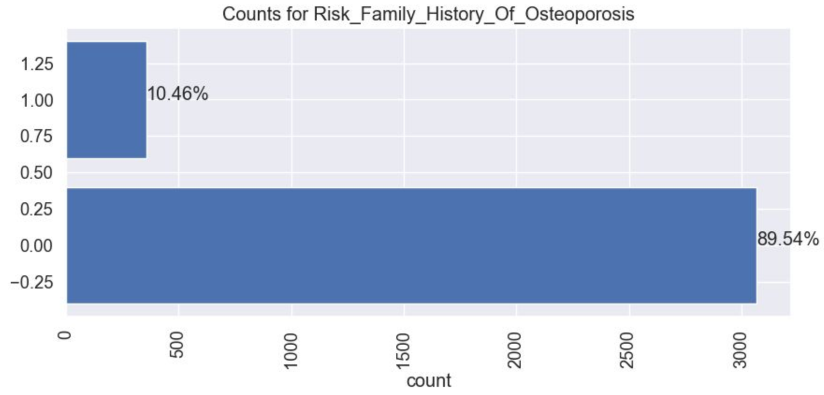
Risks

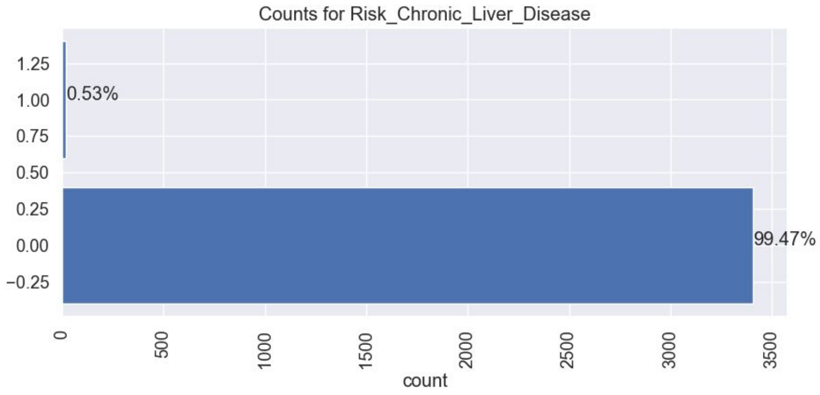




For both ‘risk_chronic_malnutrition_or_malabsorption’ and ‘risk_smoking_tobacco’, an overwhelming majority of the patients came back negative for both conditions. For ‘risk_chronic_malnutrition_or_malabsorption’, 13.73% were positive and for ‘risk_smoking_tobacco’, 18.81% were positive.

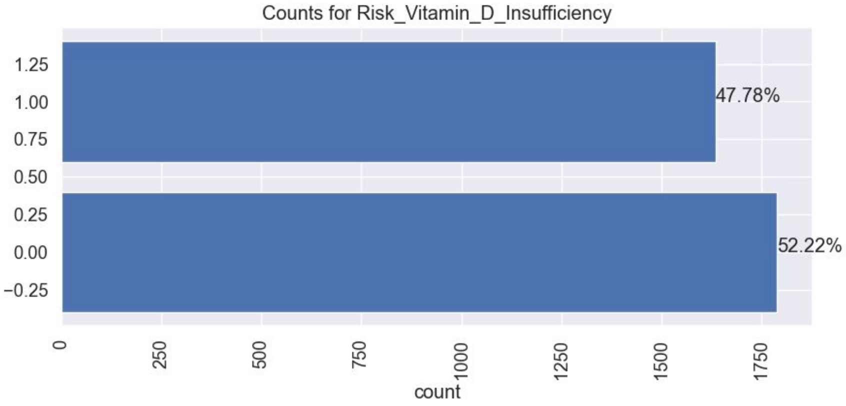
Risks

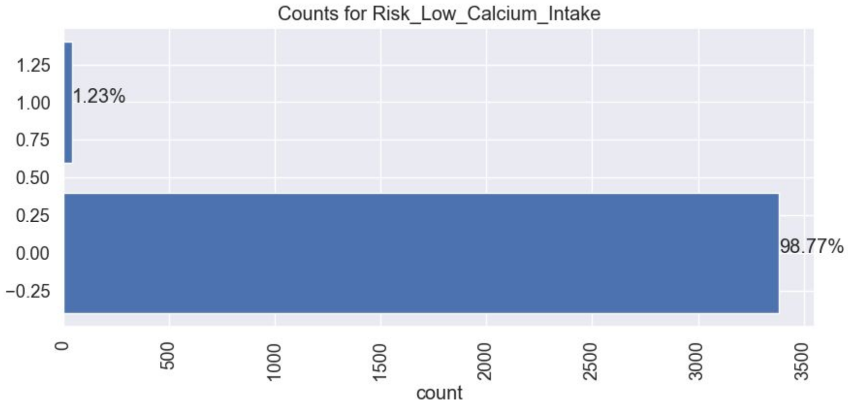




For ‘risk_family_history_of_osteoporosis’, an overwhelming majority of the patients, 89.54%, came back as negative. For ‘risk_chronic_liver_disease’, nearly all of the patients came back at negative with only 0.53% resulting in positive.

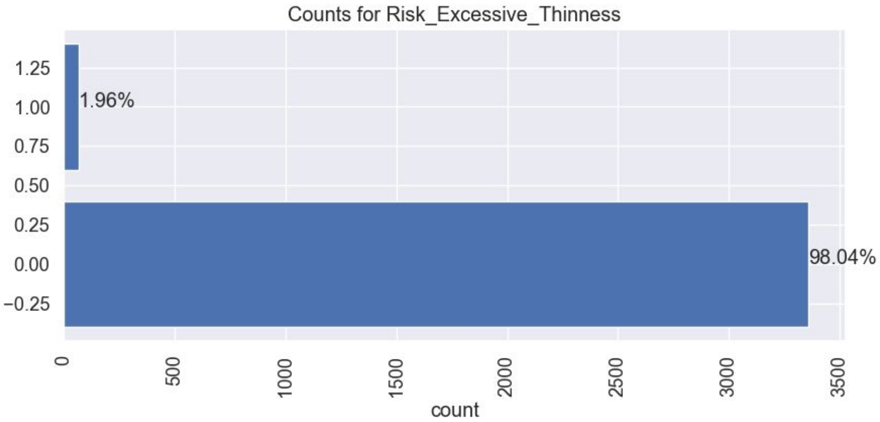
Risks

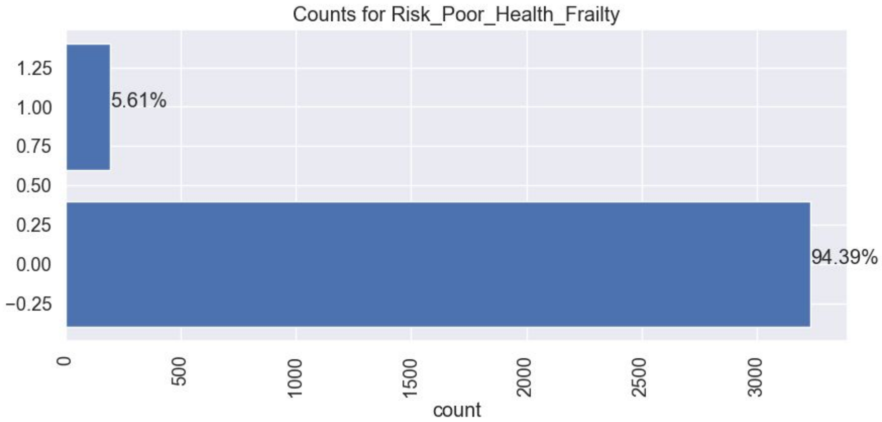




For 'risk_vitamin_D_insufficiency', about half of the patients, 52.22%, came back as negative. For 'risk_low_calcium_intake', nearly all of the patients came back at negative with only 1.23% resulting in positive.

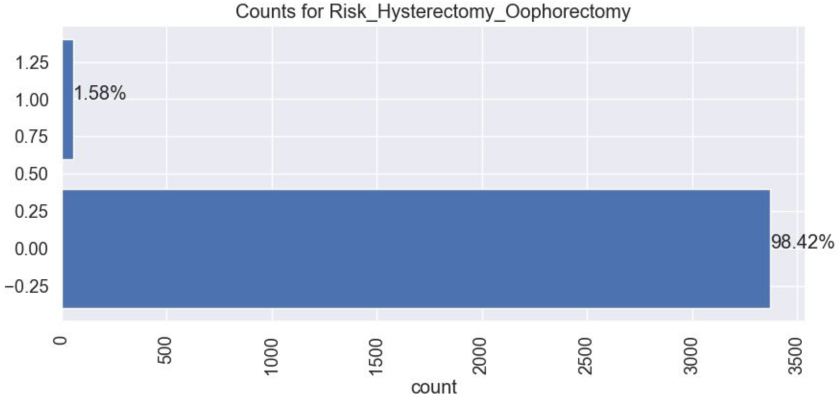
Risks

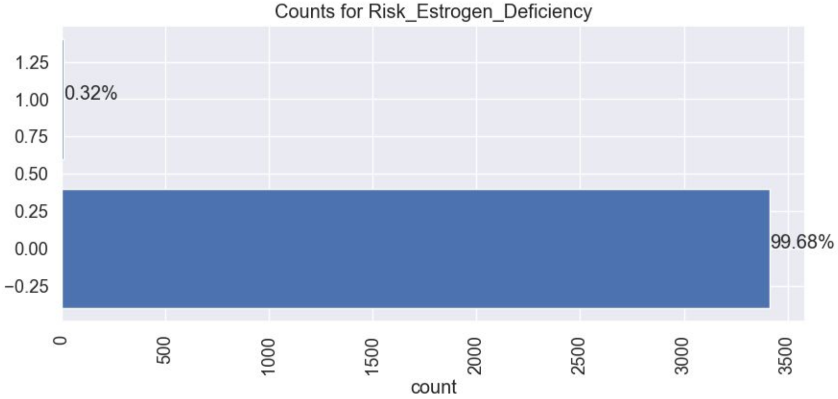




For both ‘risk_excessive_thinness’ and ‘risk_poor_health_fraility’, nearly all of the patients came back negative for both conditions. For ‘risk_excessive_thinness’, 1.96% came back as positive and for ‘risk_poor_health_fraility’ 5.61% came back as positive.

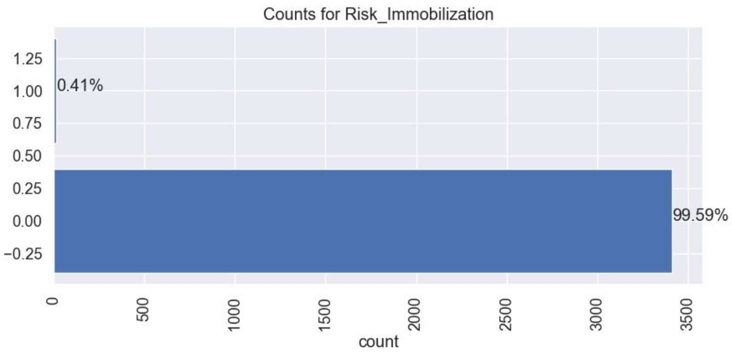
Risks





For both ‘risk_hysterectomy_oophorectomy’ and ‘risk_estrogen_defficiency’, nearly all of the patients came back negative for both conditions.

Risks



For ‘risk_immobilization’, nearly all patients came back negative for this condition, with just 0.41% positive.

Recommendations

Our recommendation moving forward to be implement a model known as **logistic regression**. It is a type of linear model that is used for binary classification. It predicts output which is a categorical dependent variable. It can successfully use predictors such as 0 or 1, yes or no, etc.