Data Science Project

Healthcare - Persistency of a drug

Group Name: BetterHealth Analytics	
Member Details	Name: Enias Vontas Country: Greece Email: vondas100@gmail.com Specialization: Data Science

Problem Description

One of the challenges for all Pharmaceutical companies is to understand the persistence of a drug as per the physician's prescription. To solve this problem ABC pharma company approached an analytics company to automate this process of identification. We have been provided with an Excel file containing some of the company's recorded data. The particular affliction that patients in this dataset were treated for is Nontuberculous Mycobacterial (NTM), which originates from a family of common organisms found in water and soil. This type of infection is rare and can affect people with damaged lungs, or with a weakened immune system. If diagnosed, a patient might need up to two years of treatment and could get infected again in the future.

The dataset contains the target variable 'Persistency_Flag' which indicates whether a patient was persistent with their medication or not. We would like to better understand the factors affecting this variable (our dependent variable). In order to do this, we have been provided with 67 other variables (our independent variables) which can be grouped in four buckets:

- Demographics: with variables such as Age, Race, Gender, etc for each patient.
- Provider Attributes: some information about the provider that wrote the prescription to the patient, with variables such as the Specialty of the Physician, a T-Score which is the result of a scan done to patients of this disease, etc.
- Clinical Factors: certain physiological attributes which could be associated with the disease, with variables such as Usage of Glucocorticoids, Frequency of a Dexa Scan etc.
- Disease/Treatment: Comorbidity factor, divided into two categories – Acute and Chronic and Concomitancy factor, i.e. concomitant drugs recorded prior to starting with the therapy.

All of the above parameters will be considered in our Machine Learning approach in order to better understand the factors affecting a patient's Medication Persistence and to more accurately classify a patient to one of the two categories of our 'Persistency_Flag' variable.

Business Understanding

It can be clear to imagine why a patient not receiving the whole dosage regimen that was prescribed to them can have unwanted results toward the treatment of that patient's illness. Another important unwanted result from this scenario is all the prescribed medication that goes to waste, from manufacturing it, all the way to distributing it to local pharmacies. So it is very important for drug companies and healthcare systems to provide the required medication to patients, but also as important for patients to be consistent with that prescribed medication, otherwise all that drug availability and expenditure would have been for nothing.

Before we talk more about our project from a business understanding, we would like to offer two definitions for a better overall understanding [**1**].

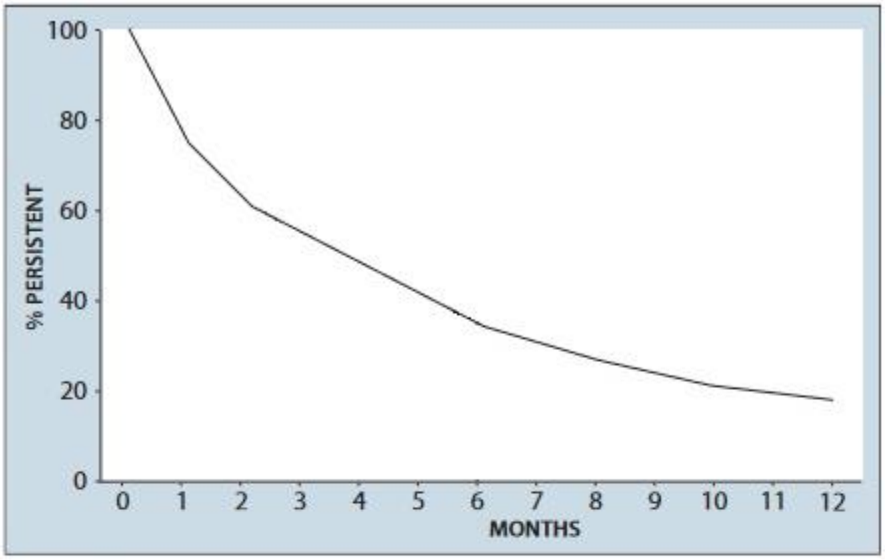
Medication Compliance (Adherence): refers to the degree of extent of conformity to the recommendation about day-to-day treatment by the provider with respect to timing, dosage, and frequency, or the extent to which a patient acts in accordance with the prescribed interval, and dose of a dosing regimen.

Medication persistence: refers to the act of continuing the treatment for the prescribed duration, or the duration of time from initiation to discontinuation of therapy.

Inadequate medication compliance and persistence are age-old problems in the pharmaceutical business. When taken in varying degrees of deviation from the prescribed dosing regimen, medications have situation-specific alterations in benefit-risk ratios, either because of reduced benefits, increased risks, or both. Numerous studies have demonstrated that inadequate compliance and non persistence with prescribed medication regimens result in increased morbidity and mortality from a wide variety of illnesses, as well as increased healthcare costs. Factoring in actual compliance and persistence is central to an accurate assessment of effectiveness and cost-effectiveness of therapy.

This source of wasted US healthcare spending every year has the potential to reach even $300 billion, while also affecting pharmaceutical companies [**2**][**3**]. A lot of factors could contribute to a patient stopping, or altering their medication regimen, they could be a physician's time constraint, competing priorities for patients and shortcoming in follow-up initiatives. These factors need to be determined by healthcare providers, as well as pharmaceutical companies in order to address them and control them as much as possible.

It is known that the drug persistence curve has a downward trend and it tends to decrease at a decreasing rate as can be seen in the figure below, where we consider the drug persistence as a percentage, and observe in the duration of a year [4]:



We would like to determine the factors affecting a patient's persistence to their prescribed medication so that companies and doctors could then control for those factors when prescribing medication.

Project Lifecycle and Deadline

The project is due on 15th of August. It has been broken into various sections,which will be completed consecutively, as presented below:

- Problem Understanding
- Business Understanding
- Data Understanding
- Data Cleaning and Feature Engineering
- Model Development
- Model Selection
- Model Evaluation
- Report the accuracy, precision and recall of both the classes of target variable
- Report ROC-AUC
- Deploy the model
- Explain Challenges and Model Selection

As a first section, we will focus on the first two points, that are also underlined. As the project progresses, we will move forward with the other sections, as well as re-evaluate our findings if needed.

Data Understanding

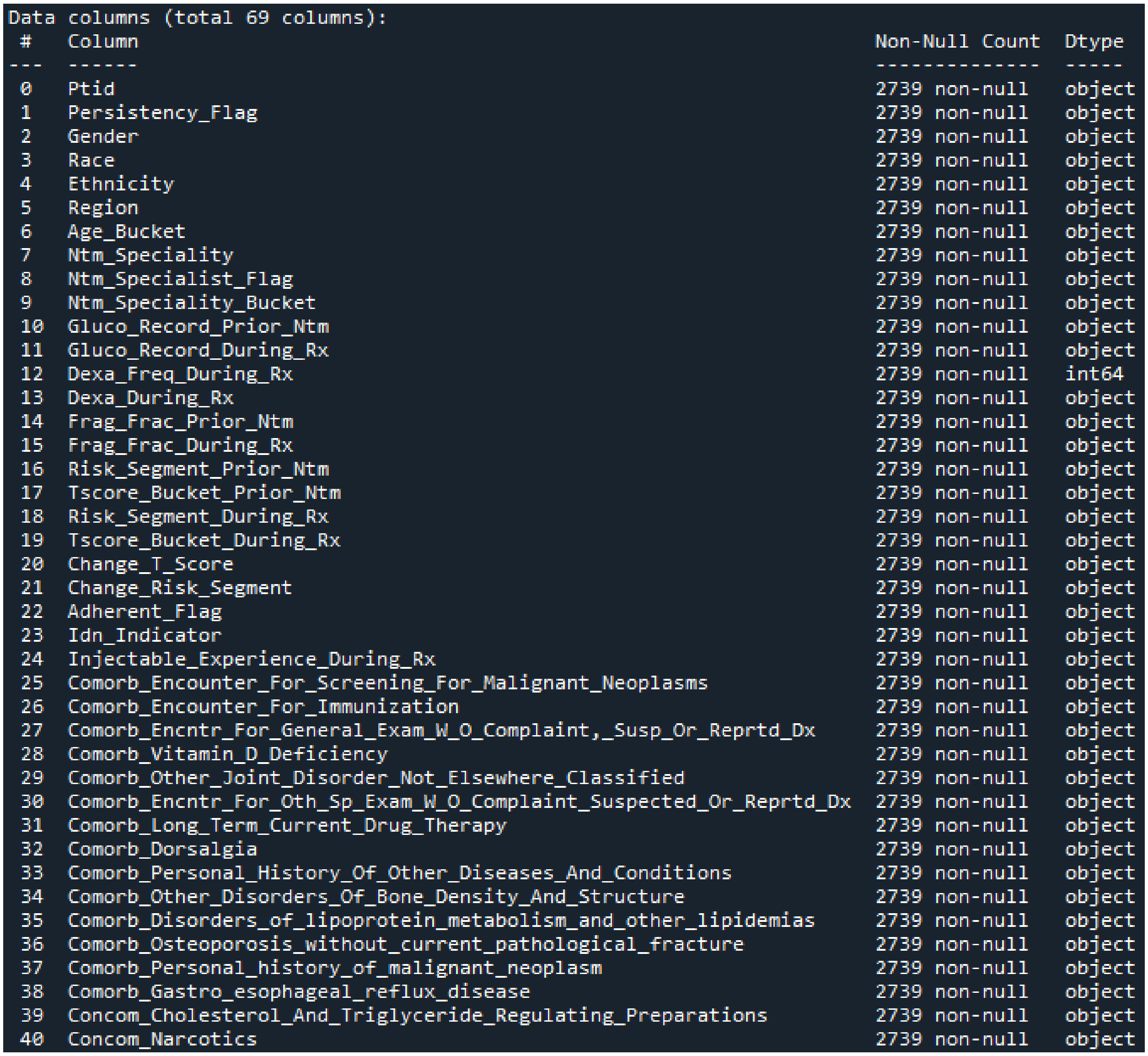
The dataset provided to us contains 3424 patients (each with their own ID), our target variable, the Persistency of the drug, and 67 other variables, which we will use as predictors for our target/dependent variable. The predictor variables can be grouped in four different buckets: 'Demographics', 'Provider Attributes', 'Clinical Factors' and 'Disease/Treatment Factors'. Almost all our predictor variables are categorical and ordinal with the exception of two ('Dexa Scan Frequency' and 'Count of Risks') which are numerical. Our ordinal variables ('Age_Bucket' and T-Scores) are 4 in total and the rest are categorical, either with two categories ('Yes' or 'No') or with more than two categories ('No Change', 'Unkown', 'Worsened', 'Improved').

We provide a table below with all the variables and a brief description for each one:

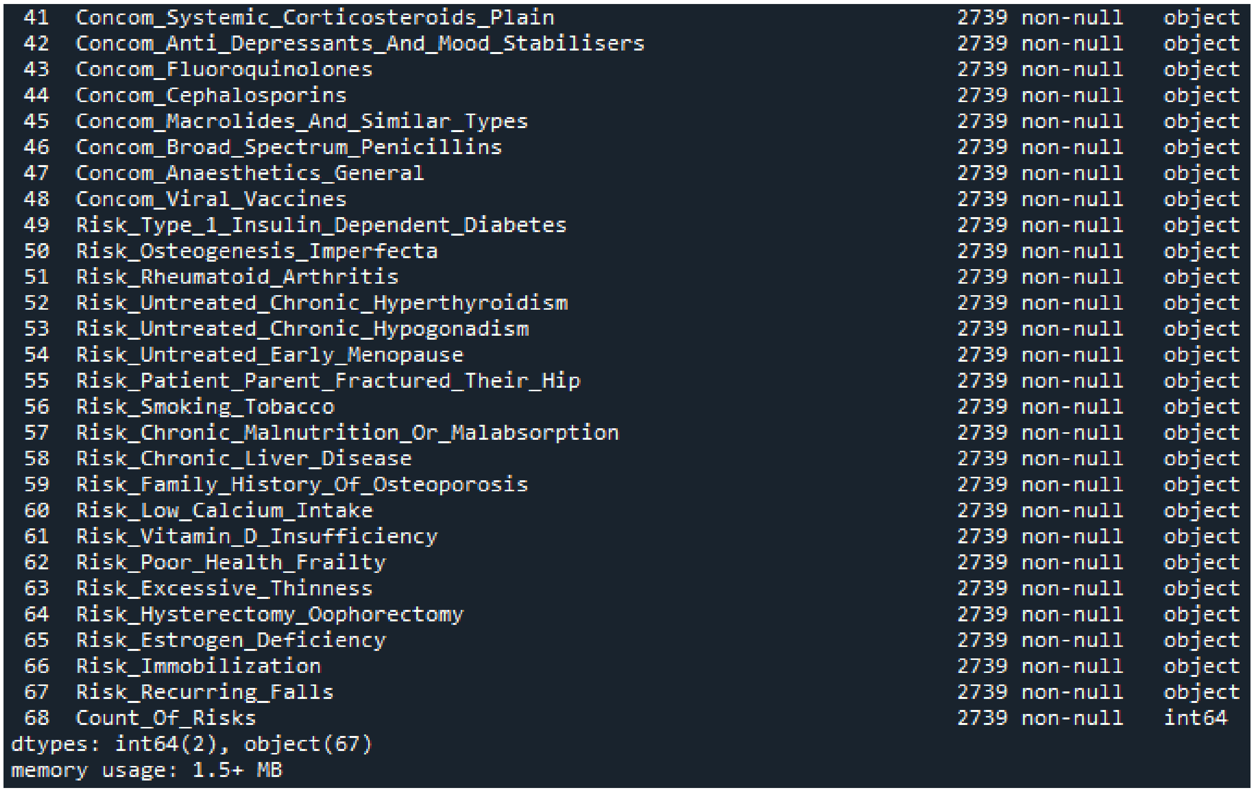
Bucket	**Variable**	**Variable Description**
Unique Row Id	*Patient ID*	Unique ID of each patient
Target Variable	*Persistency_Flag*	Flag indicating if a patient was persistent or not
Demographics	*Age*	Age of the patient during their therapy
	Race	Race of the patient from the patient table
	Region	Region of the patient from the patient table
	Ethnicity	Ethnicity of the patient from the patient table
	Gender	Gender of the patient from the patient table
	IDN Indicator	Flag indicating patients mapped to Integrated Deliver Network
Provider Attributes	*NTM - Physician Specialty*	Specialty of the Health Care Personnel that prescribed the NTM Rx
Clinical Factors	*NTM - T-Score*	T Score of the patient at the time of the NTM Rx (within 2 years prior from rxdate)
	Change in T- Score	Change in Tscore before starting with any therapy and after receiving therapy
	NTM - Risk Segment	Risk Segment of the patient at the time of the NTM Rx (within 2 years prior to rxdate)
	Change in Risk Segment	Change in Risk Segment before starting any therapy and after receiving therapy
	NTM - Multiple Risk Factors	Flag indicating if patient falls under multiple risk category at the time of the NTM Rx (within 365 days prior to rxdate)
	NTM - Dexa Scan Frequency	Number of DEXA scans taken prior to the first NTM Rx(within 365 days prior to rxdate)
	NTM - Dexa Scan Recency	Flag indicating the presence of Dexa Scan before the NTM Rx (within 2 years prior from rxdate or between their first Rx and Switched Rx; whichever is smaller and applicable)
	Dexa During Therapy	Flag indicating if the patient had a Dexa Scan during their first continuous therapy
	NTM - Fragility Fracture Recency	Flag indicating if the patient had a recent fragility fracture (within 365 days prior from rxdate)
	Fragility Fracture During Therapy	Flag indicating if the patient had fragility fracture during their first continuous therapy
	NTM - Glucocorticoid Recency	Flag indicating usage of Glucocorticoids (>=7.5mg strength) in the one year look-back from the first NTM Rx
	Glucocorticoid Usage During Therapy	Flag indicating if the patient had a Glucocorticoid usage during the first continuous therapy
Disease/Treatment Factors	*NTM - Injectable Experience*	Flag indicating any injectable drug usage in the recent 12 months before the NTM OP Rx
	NTM - Risk Factors	Risk Factors that the patient is falling into. For chronic Risk Factors complete lookback to be applied and for non-chronic Risk Factors, one year lookback from the date of first OP Rx
	NTM - Comorbidity	Comorbidities are divided into two main categories - Acute and chronic, based on the ICD codes. For chronic disease we are taking complete look back from the first Rx date of NTM therapy and for acute diseases, time period before the NTM OP Rx with one year lookback has been applied
	NTM - Concomitancy	Concomitant drugs recorded prior to starting with a therapy(within 365 days prior from first rxdate)
	Adherence	Adherence for the therapies

It is very important to note, that since we have been assigned a Classification Machine Learning problem, we will split our dataset into a ‘Train’ and a ‘Test’ set with 80% and 20% of the patients respectively. This is being done in order to ‘Train’ our model first and then evaluate its performance on the ‘Test’ set, which we consider as unknown at this point, so as not to ‘contaminate’ our model building process. There are different schools of thought as to when this split should be done, but **from this point forward** all analysis is being done on the ‘Train’ set, unless specified otherwise.

As far as any possbible NA or missing values are concerned, we did not find any:



```
Data columns (total 69 columns):
 #   Column                                                              Non-Null Count  Dtype
---  ------                                                              --------------  -----
 0   Ptid                                                                2739 non-null   object
 1   Persistency_Flag                                                    2739 non-null   object
 2   Gender                                                              2739 non-null   object
 3   Race                                                                2739 non-null   object
 4   Ethnicity                                                           2739 non-null   object
 5   Region                                                              2739 non-null   object
 6   Age_Bucket                                                          2739 non-null   object
 7   Ntm_Speciality                                                      2739 non-null   object
 8   Ntm_Specialist_Flag                                                 2739 non-null   object
 9   Ntm_Speciality_Bucket                                               2739 non-null   object
 10  Gluco_Record_Prior_Ntm                                              2739 non-null   object
 11  Gluco_Record_During_Rx                                              2739 non-null   object
 12  Dexa_Freq_During_Rx                                                 2739 non-null   int64
 13  Dexa_During_Rx                                                      2739 non-null   object
 14  Frag_Frac_Prior_Ntm                                                 2739 non-null   object
 15  Frag_Frac_During_Rx                                                 2739 non-null   object
 16  Risk_Segment_Prior_Ntm                                              2739 non-null   object
 17  Tscore_Bucket_Prior_Ntm                                             2739 non-null   object
 18  Risk_Segment_During_Rx                                              2739 non-null   object
 19  Tscore_Bucket_During_Rx                                             2739 non-null   object
 20  Change_T_Score                                                      2739 non-null   object
 21  Change_Risk_Segment                                                 2739 non-null   object
 22  Adherent_Flag                                                       2739 non-null   object
 23  Idn_Indicator                                                       2739 non-null   object
 24  Injectable_Experience_During_Rx                                     2739 non-null   object
 25  Comorb_Encounter_For_Screening_For_Malignant_Neoplasms              2739 non-null   object
 26  Comorb_Encounter_For_Immunization                                   2739 non-null   object
 27  Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx     2739 non-null   object
 28  Comorb_Vitamin_D_Deficiency                                         2739 non-null   object
 29  Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified                2739 non-null   object
 30  Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx  2739 non-null   object
 31  Comorb_Long_Term_Current_Drug_Therapy                               2739 non-null   object
 32  Comorb_Dorsalgia                                                    2739 non-null   object
 33  Comorb_Personal_History_Of_Other_Diseases_And_Conditions            2739 non-null   object
 34  Comorb_Other_Disorders_Of_Bone_Density_And_Structure                2739 non-null   object
 35  Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias     2739 non-null   object
 36  Comorb_Osteoporosis_without_current_pathological_fracture           2739 non-null   object
 37  Comorb_Personal_history_of_malignant_neoplasm                       2739 non-null   object
 38  Comorb_Gastro_esophageal_reflux_disease                             2739 non-null   object
 39  Concom_Cholesterol_And_Triglyceride_Regulating_Preparations         2739 non-null   object
 40  Concom_Narcotics                                                    2739 non-null   object
```



```
 41   Concom_Systemic_Corticosteroids_Plain                                 2739 non-null   object
 42   Concom_Anti_Depressants_And_Mood_Stabilisers                          2739 non-null   object
 43   Concom_Fluoroquinolones                                               2739 non-null   object
 44   Concom_Cephalosporins                                                 2739 non-null   object
 45   Concom_Macrolides_And_Similar_Types                                   2739 non-null   object
 46   Concom_Broad_Spectrum_Penicillins                                     2739 non-null   object
 47   Concom_Anaesthetics_General                                           2739 non-null   object
 48   Concom_Viral_Vaccines                                                 2739 non-null   object
 49   Risk_Type_1_Insulin_Dependent_Diabetes                                2739 non-null   object
 50   Risk_Osteogenesis_Imperfecta                                          2739 non-null   object
 51   Risk_Rheumatoid_Arthritis                                             2739 non-null   object
 52   Risk_Untreated_Chronic_Hyperthyroidism                                2739 non-null   object
 53   Risk_Untreated_Chronic_Hypogonadism                                   2739 non-null   object
 54   Risk_Untreated_Early_Menopause                                        2739 non-null   object
 55   Risk_Patient_Parent_Fractured_Their_Hip                               2739 non-null   object
 56   Risk_Smoking_Tobacco                                                  2739 non-null   object
 57   Risk_Chronic_Malnutrition_Or_Malabsorption                            2739 non-null   object
 58   Risk_Chronic_Liver_Disease                                            2739 non-null   object
 59   Risk_Family_History_Of_Osteoporosis                                   2739 non-null   object
 60   Risk_Low_Calcium_Intake                                               2739 non-null   object
 61   Risk_Vitamin_D_Insufficiency                                          2739 non-null   object
 62   Risk_Poor_Health_Frailty                                              2739 non-null   object
 63   Risk_Excessive_Thinness                                               2739 non-null   object
 64   Risk_Hysterectomy_Oophorectomy                                        2739 non-null   object
 65   Risk_Estrogen_Deficiency                                              2739 non-null   object
 66   Risk_Immobilization                                                   2739 non-null   object
 67   Risk_Recurring_Falls                                                  2739 non-null   object
 68   Count_Of_Risks                                                        2739 non-null   int64
dtypes: int64(2), object(67)
memory usage: 1.5+ MB
```

We can see that for each variable we have 2739 non-null values, meaning that there are no null data points in our dataset. But there are 5 variables: Ntm_Speciality, T score During Rx, Change in T score, Risk Segment During Rx and Change in Risk Segment which have an 'Unknown' category, meaning that no value was observed. In the table below we have the number of 'Unknown' counts in each of them:

Variable	**'Unknown' category counts**
Ntm_Speciality	258/2739 = 9.42%
Risk_Segment_During_Rx	1223/2739 = 44.65%
Change_Risk_Segment	1802/2739 = 65.8%
Tscore_Bucket_During_Rx	1223/2739 = 44.65%
Change_T_Score	1223/2739 = 44.65%

The variable 'Change in Risk Segment' has over 65% of its observations marked as 'Unknown', so we decide to drop this column altogether. As for the other variables, we will try to Impute the missing values, since they are less than 60% in each column. The Imputation method decided for each variable will be performed on Training and Test sets so as not to 'leak' information in our training set, which we will use for model building, from our test set, which we will use for classification of the Persistency Flag variable.

Imputation

Rubin [**5**] classified missing data problems into three categories. In his theory every data point has some likelihood of being missing.

1. If the probability of being missing is the same for all cases, then the data are said to be missing completely at random (**MCAR**), meaning that if a certain value is missing, it has nothing to do with hypothetical value and with the values of other variables.
2. If the probability of being missing is the same only within groups defined by the *observed* data, then the data are missing at random (**MAR**), meaning that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data.
3. If the probability of being missing varies for reasons that are unknown to us, then the data are said to be missing not at random (**MNAR**), meaning that missing value depends on a hypothetical value, or on some other variable's value. Usual strategy for this case is to gather more data, which in our case we cannot do at the moment.

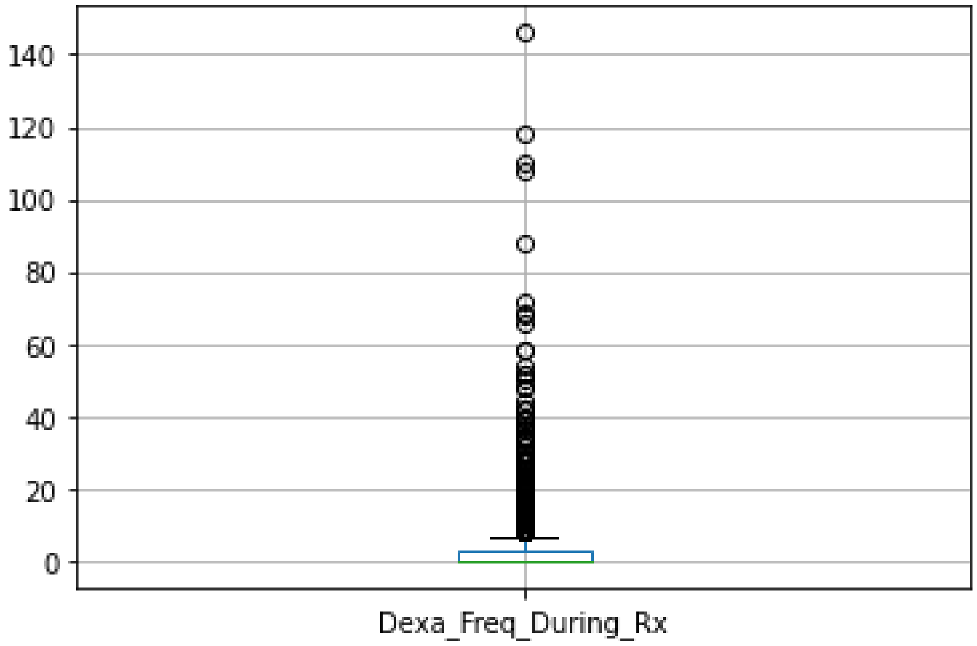
We will consider our variables' missing values to be unrelated to the missing data, but could be related to the observed data (MAR), so we will impute them. For all these variables, we will impute the missing values with the KNNImputer method, with 'number of neighbors' = 1. This method takes into account the whole dataset and each missing feature is imputed using values from its nearest neighbor, where distance is calculated via a eucledian metric that supports missing values (**nan_euclidean_distances**[**6**]). We present below as an example two of the variables in our training set, before and after imputation:

Variable	**Before Imputation**	**After Imputation**
Risk_Segment_During_Rx	High Risk: 763 Unknown: 1421 Low Risk: 753 Total: 2937	High Risk: 1292 Low Risk: 1645 Total: 2937
Tscore_Bucket_During_Rx	<=-2.5: 805 >-2.5: 711 Unknown : 1421 Total: 2937	<=-2.5: 1491 >-2.5: 1446 Total: 2937

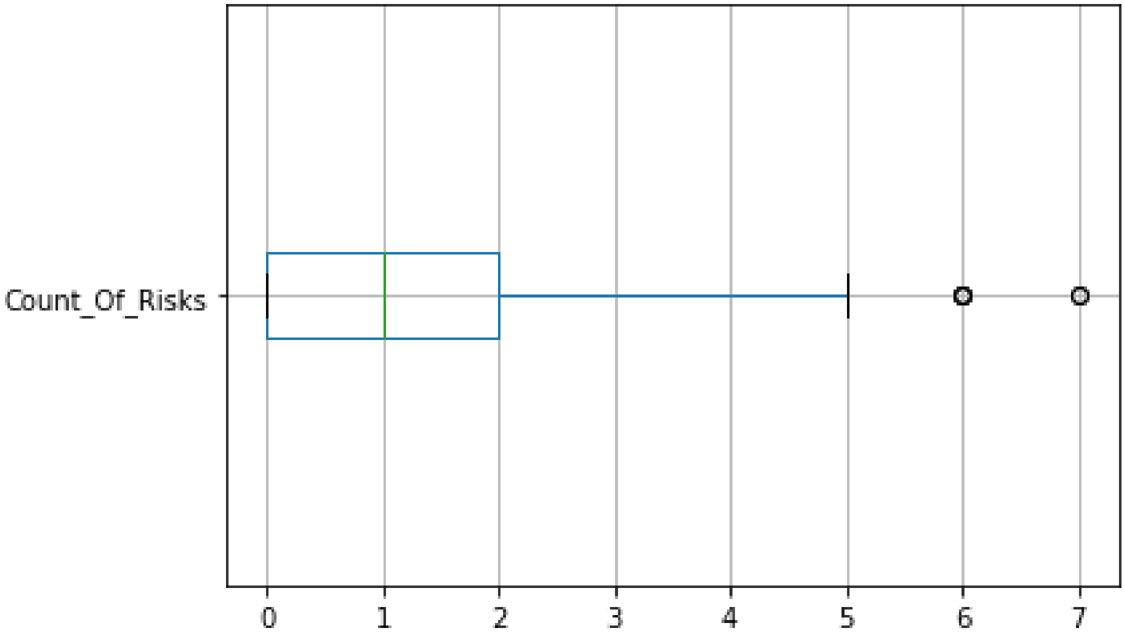
There are no more missing values, 'Unknown' label, and the values imputed do not seem to have affected the distribution of the labels by much.

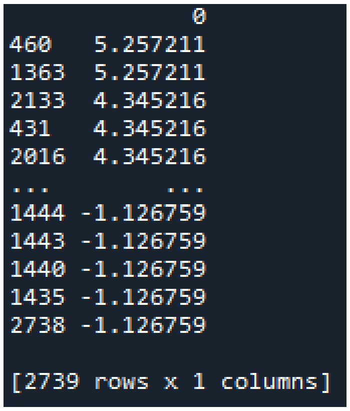
Outliers

We present below the boxplot for the Frequency of Dexa Scans the patients did, where we have the number of scans taken during the patient's first continuous therapy. An NTM infection therapy can take many months. A patient is considered cured when samples taken from them show no sign of NTM infection for at least 12 months [**7**]. A big percentage of the patients did not have a Dexa Scan (2488, or 72.66%), as can be seen by the distribution plotted in the boxplot, which has a mean of 3 scans and 75% of our observation (3rd Quatrile) are up to 3 scans. For the rest of the 25% of our observations, we can observe some values that can be considered quite high, some even greater than 100, meaning that in a 36 month therapy period the patient had a Dexa scan every 7.5 days. For patients with osteoporosis, one scan every year is recommended [**8**] so there are numbers presented here that mgiht seem out of the ordinary. Unfortunately, we do not have further information about these patients, whether they actually did that many scans or these values might be typos, or miscalculations. And so, we cannot remove these values, just on the basis that they are inconvenient.



A different picture can be seen in our ‘Count of Risks’ variable, as presented in the figure below. A median value of 1, with 75% of the patients presenting from 0 up to 5 possible Risk factors, and 2 patients having 7, and 6 patients having 6 Risk factors. We also have the Z-scores of this variable, where we can see that the 2 patients with 7 Risk factors are 5.26 Standard Deviations away from the variable’s mean, while the other 6 patients with 6 Risk factors are 4.34 Stadard Deviations away from the variable’s mean. While these are values that are very far away from the mean, they cannot be easily considered as outliers, and so we do not remove them from our analysis.



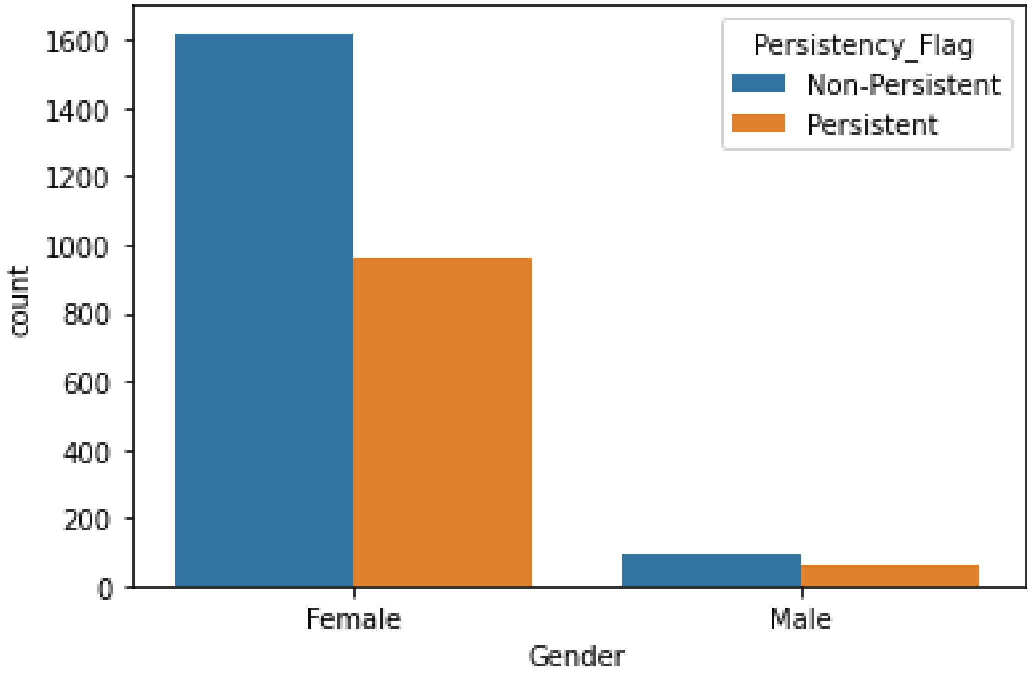

```
             0
460   5.257211
1363  5.257211
2133  4.345216
431   4.345216
2016  4.345216
...        ...
1444 -1.126759
1443 -1.126759
1440 -1.126759
1435 -1.126759
2738 -1.126759

[2739 rows x 1 columns]
```

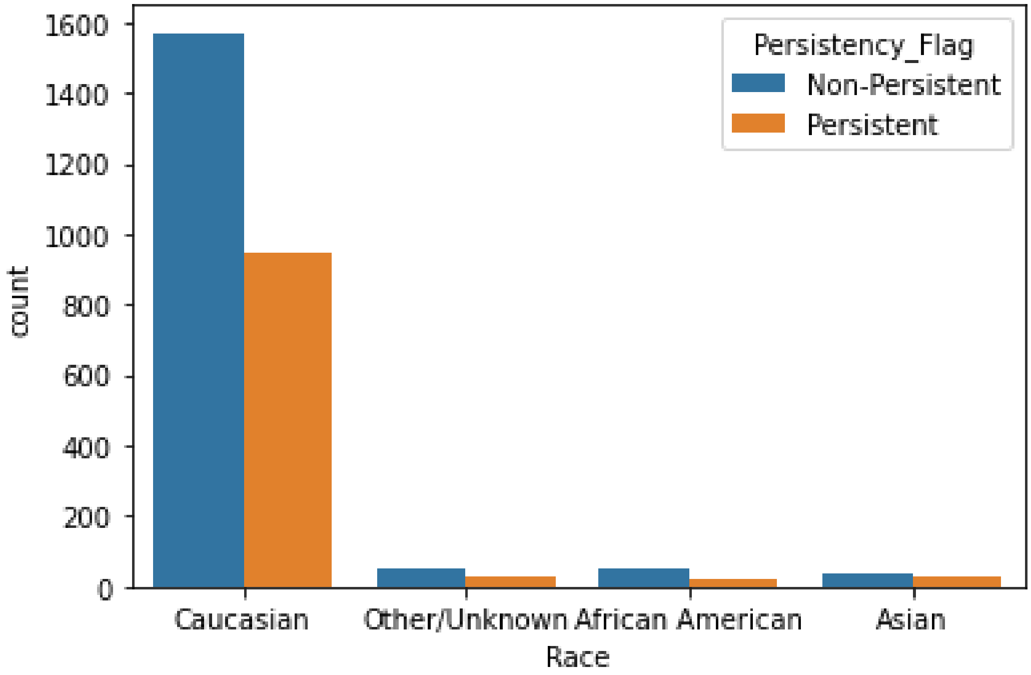
Exploratory Data Analysis

We will start with the **Demographics** Bucket:

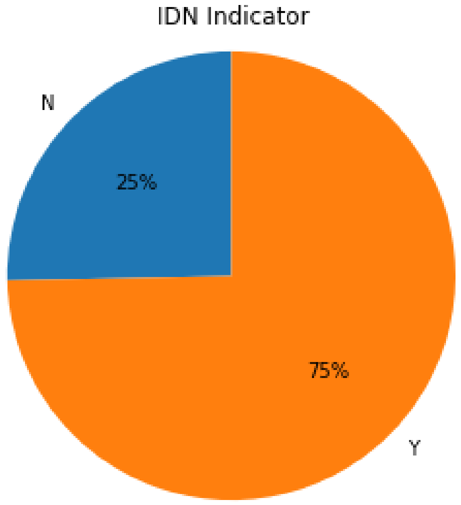
We have below the count plot for the gender of our patients, where 2582 (94.3%) are Female and 157 (5.73%) are Male. For the Female patients, 1620 (62.74%) were non-persistent and 962 (37.36%) were Persistent, while for Male patients, 93 (59.24%) were non-persistent and 64 (40.76%) were Persistent with their prescriptions. So we have a similar picture in both Genders, despite the fact that the vast majority of our patients are Female.



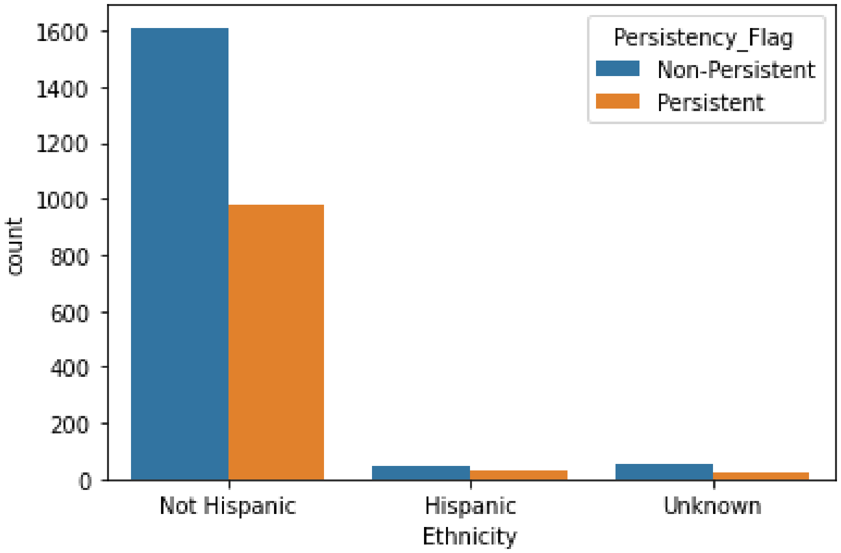
Presented below is the count plot for the Races of our patients and again grouped by their Persistency. We can see that most of our patients were Caucasian (2513, or 91.75%), followed by 'Other/Unknown', then African Americans and then Asians. Again we can see that in all groups we have more non-persistent patients than persistent, in the Caucasian group, we have 62.5% being non-persistent and 37.5% being persistent with their prescriptions.



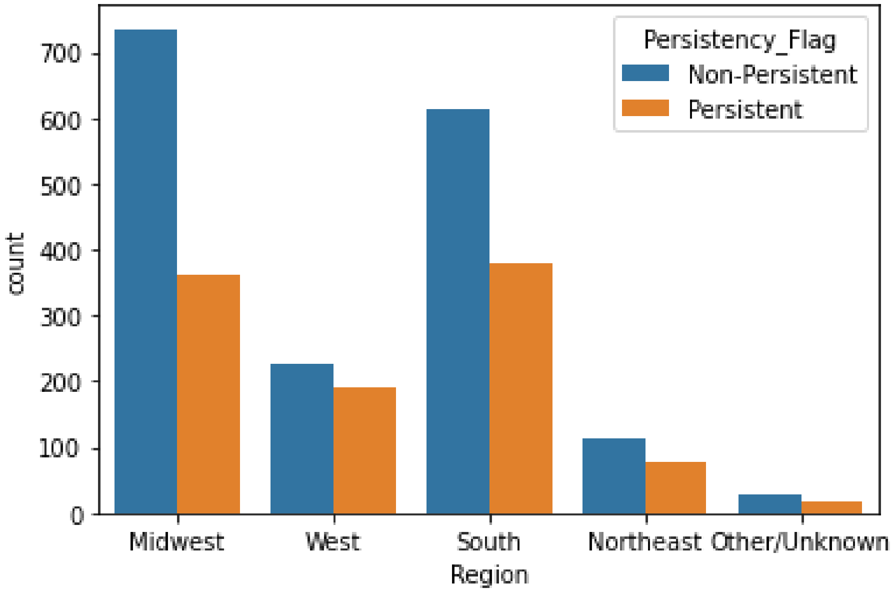
In the pie chart below we have the percentage of patients that were mapped to an **Integrated Delivery Network** (IDN). More than 76% of U.S. hospitals are part of an IDN. Pharma companies research hospitals and care facilities in each of their target IDNs in order to maximize the adoption of their drug/therapy. This is one indicator that health care industry consolidation means most of the people making purchasing decisions (such as physicians and C-Suite executives) are responsible for leading wider care networks. Our dataset seems to follow the wider population trend, with 75% of the patient being mapped to an IDN.



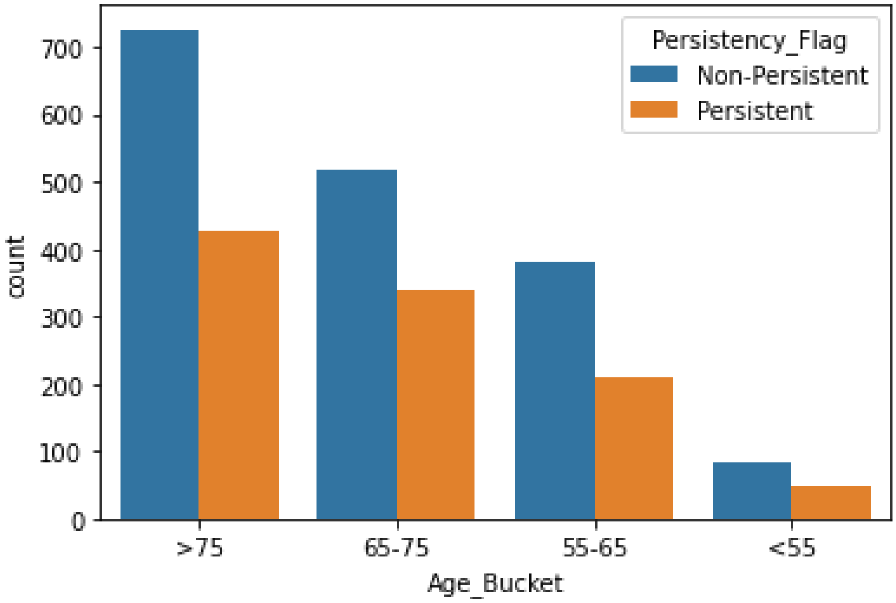
In the figure below, we have the Ethnicities of our patients, with 94.48% being non-Hispanic, 2.7% being Hispanic and 2.8% of 'Unknown' ethnicity. Again we have a picture of more people in each group being non-persistent than persistent with their prescriptions.



In the following image, we have the regions of our patients and again they have been grouped according to their drug persistency. Most people come from the Midwest (1098) closely followed by South Region (991) and then West (416), with Northeast and Unknown in the end. We can observe that patients in our regions tend to be more non-persistent than persistent.



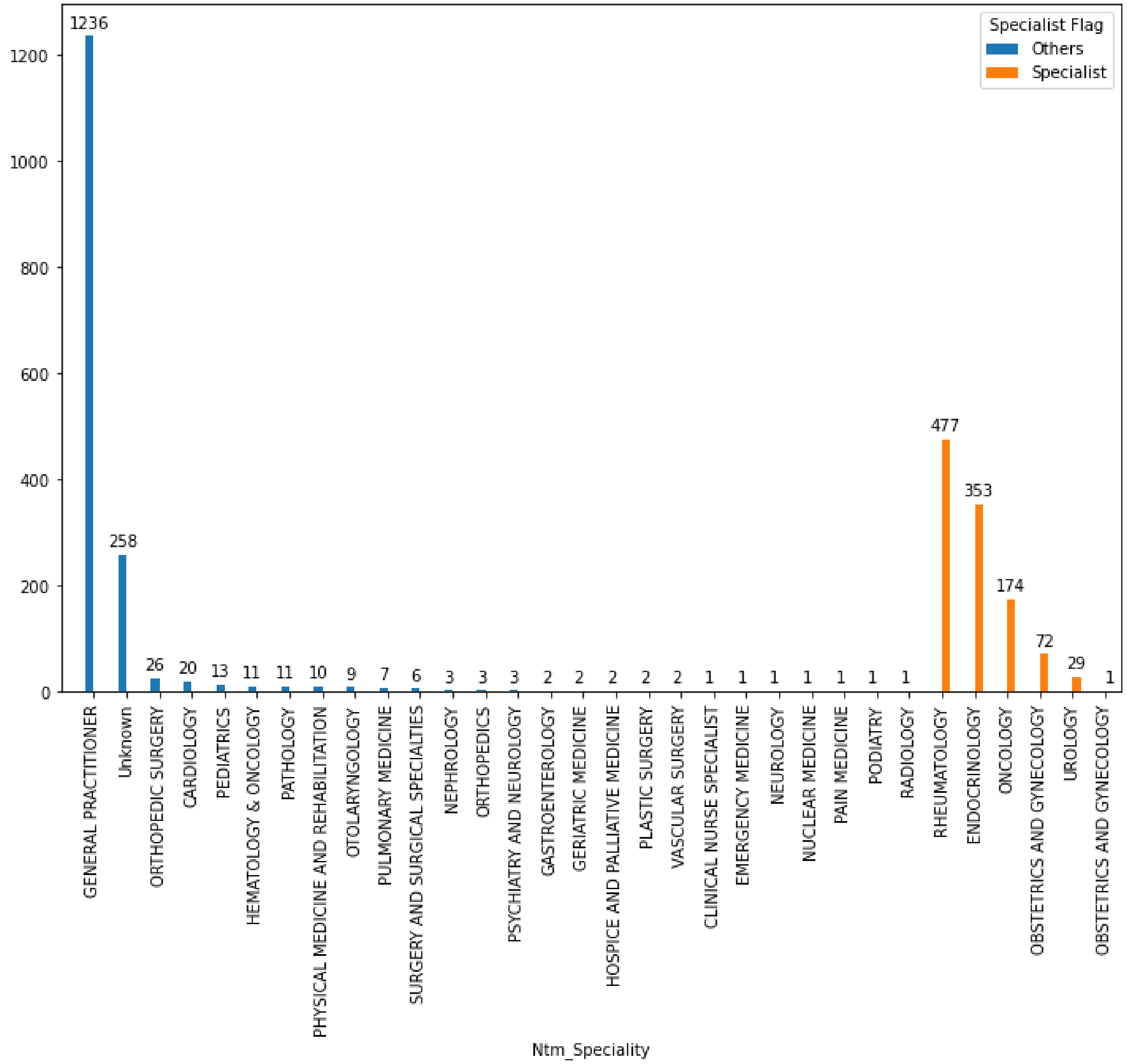
In the image presented below, we can see the age groups of our patients and the drug persistency in each group. Again more patients are non-persistent than they are persistent. We have 42.17% in the category '>75' years old, followed by 31.4% in the category '65-75', then 21.58% in the '55-65' years old category, with 4.9% in the last category '<55'.



We have seen that most our patients are non-Hispanic, Caucasian Females, in the age categories of ’65-75’ and ‘>75’ years old. They come from Midwest and South regions of the US. These characteristics for the patients seem fitting, since NonTuberculous Mycobacterial infections have predilition to affect thin, post-menopausal women without underlying lung diseases [**9**].

A tendency of people being more ‘non-persistent’ than ‘persistent’ can also be assumed, according to the graphs above. In total we have 62.54% of the 2739 patients being non-persistent and the rest 37.46% as being persistent with their prescriptions.

We will continue with the **‘Physician Attributes’** bucket. In the figure below, we can see the number of prescrition given by each category of Health Care Personnel (**HCP)**. We have grouped these categories on whether they are considered Specialists or not. There are 1106 Specialists in total and 1633 Others. We can see that a big percentage of prescriptions was given by General Practitioners (1236, or 45.13%), almost half of all our patients received their prescription from a General Practitioner. This could indicate how people are being informed on whether they have the NTM infection or not, and how they are then given treatment. As for the Specialist category, we can see that there 477 Rheumatologists and 353 Endocrinologists that make up the majority of this category. We have 258 HCP that are of unknown title. We will impute these values from the rest of the categories.

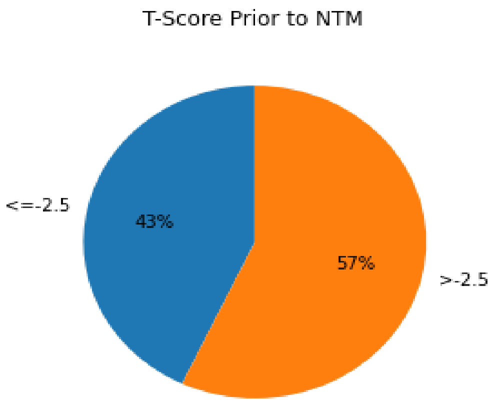


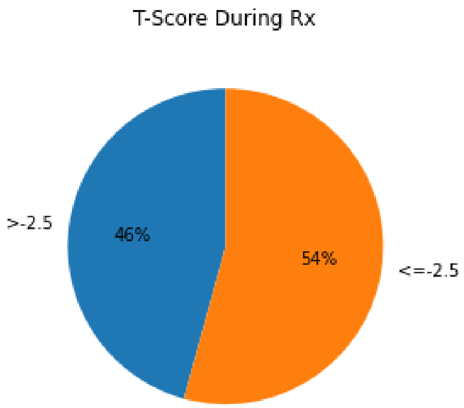
Clinical Factors

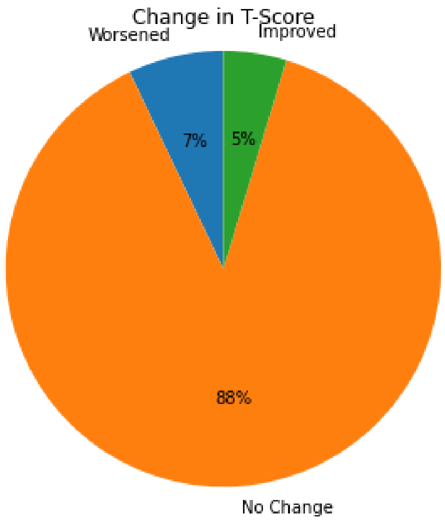
In this bucket of variables we have predictors for the Dexa scans, their results, which are calculated in T-Scores, Fragility Fractures and Glucocorticoid use. We have already seen the Dexa scan variables, so we move on to the **T-Scores** produced by these scans. Every Dexa Scan or a dual-energy X-ray absorptiometry (bone densitometry) determines bone mineral density (BMD). This BMD is compared to healthy young adults from 25-35 years old (T-Score) . The standard deviation (SD) is the difference between the patient's BMD and that of a healthy young adult:

- T-Score ±1 SD indicates normal bone density
- T-Score -1 to -2.5 SD indicates low bone mass
- T-Score <= -2.5 SD indicates the presence of osteoporosis

In general, the risk for bone fracture doubles with every standard deviation below normal.

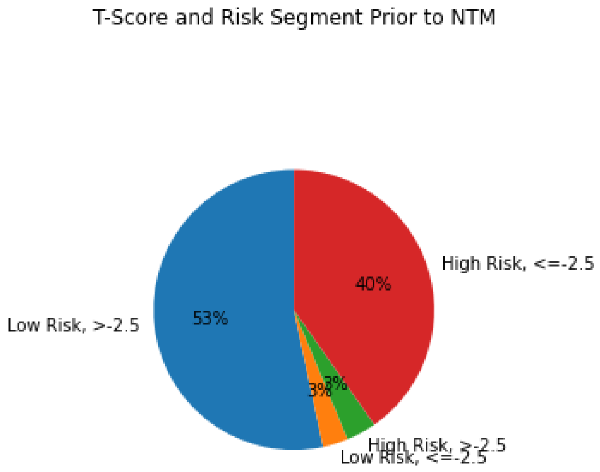


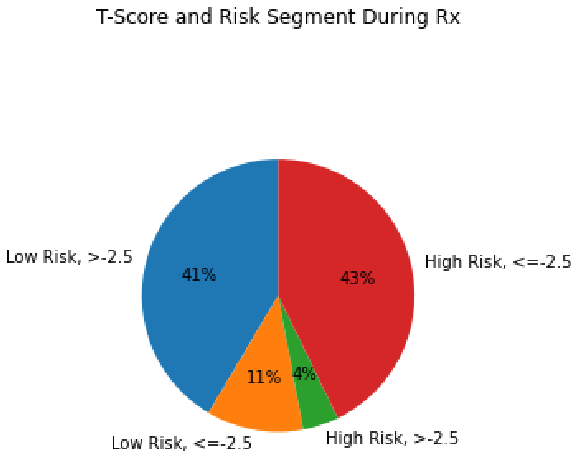




From the figures above we can see that even prior to NTM, 43% patients had a T-Score that indicates low Bone Density. After beginning therapy, we have a large percentage, 54% with low T-Scores. In the last graph, we have the change in the patients' T-Scores before starting any therapy and after receiving therapy. Around 88% show no change in their T-Scores values and only 5% showed improvement and 7% actually showed even lower T-Scores after receiving therapy for NTM infection.

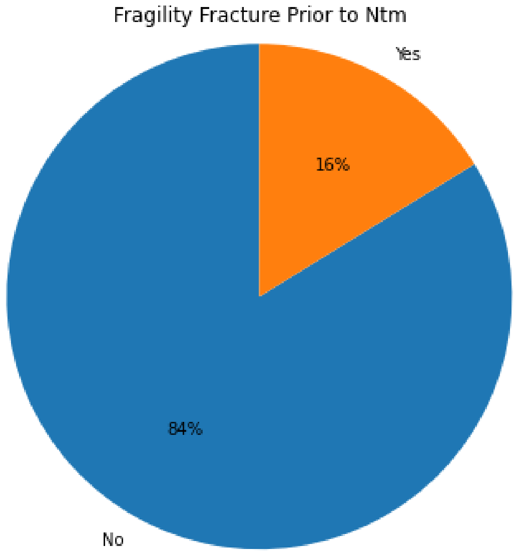
In the figures below, we have plotted the **Risk Segment** variables against the T-Scores obtained for the Dexa scans of the patients.

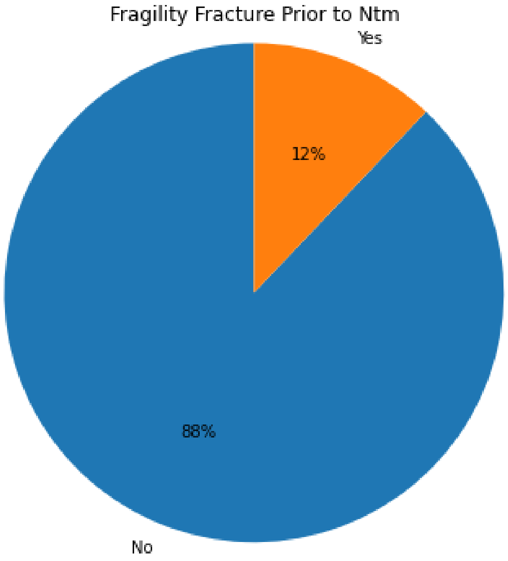




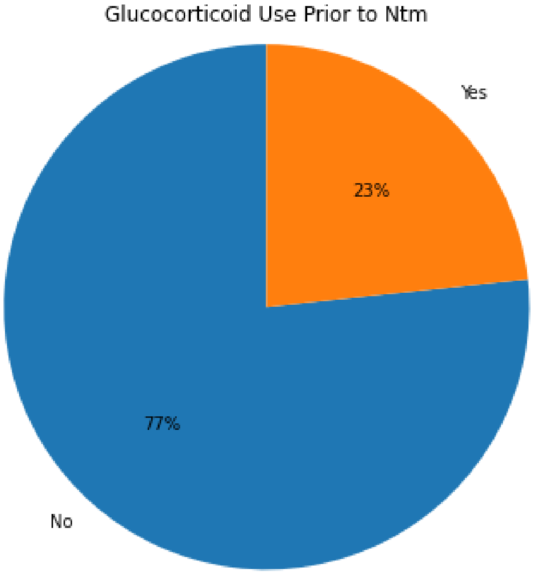
We have combinations of Risk segment and T-Scores in the first two figures. We can see that there is a tendency for these two attributes to correlate, meaning that patients considered 'High Risk' will also have a T-Score value that indicates low bone density and likewise, patients considred 'Low Risk' will have a T-Score value that indicates normal bone density.

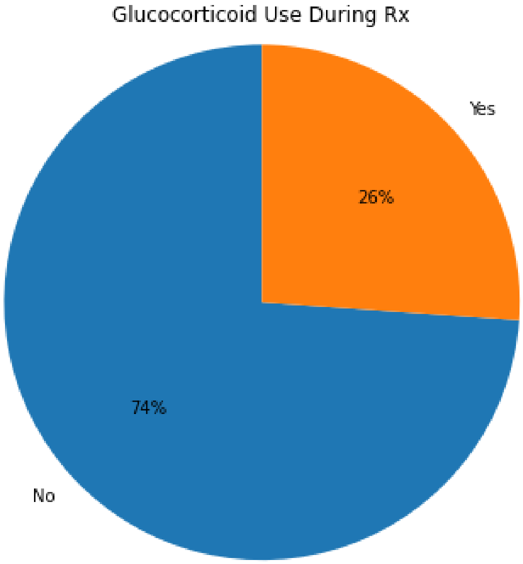
In the two pie charts below, we can see the **Fragility Fracture** predictor, which refers to fractures that result from a fall from a standing height or less. We can see that most our patients, 84% did not exhibit such a Fracture prior to being diagnosed and this percentage actually grew smaller after the patients received treatment for Ntm, since it became 88%.





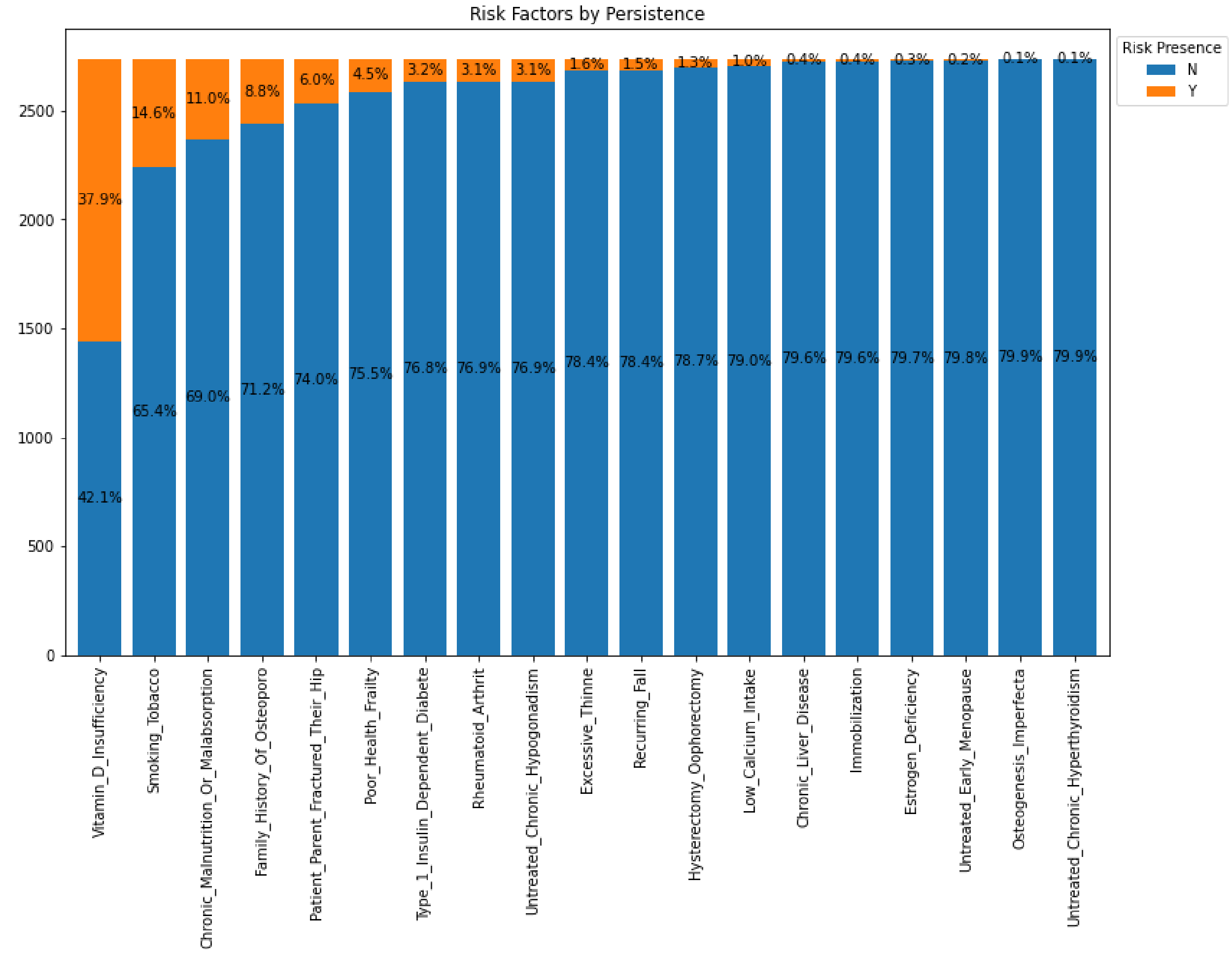
In the figures below we can see **Glucocorticoid Use** from the patients, prior and during their therapies. Glucocorticoids are powerful medicines that fight inflammation and work with the patient's immune system to treat a wide range of health problems. We can see that most patients, 77%, did not use any Glucocorticoids before their treatment and that percentage fell by just 3%, so there are 74% of patients who used Glucocorticoids along with their treament.



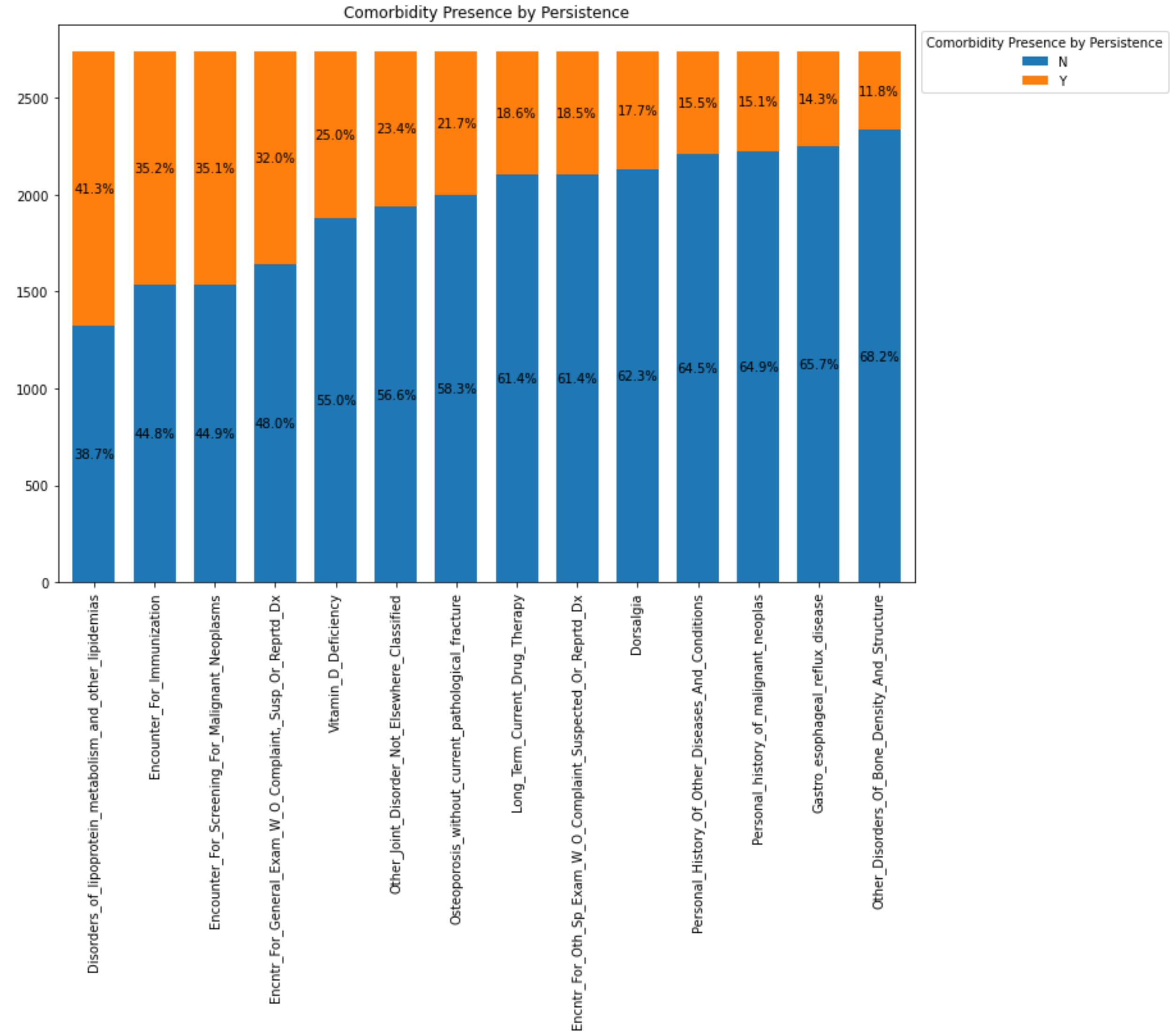


Disease/Treatment Factors

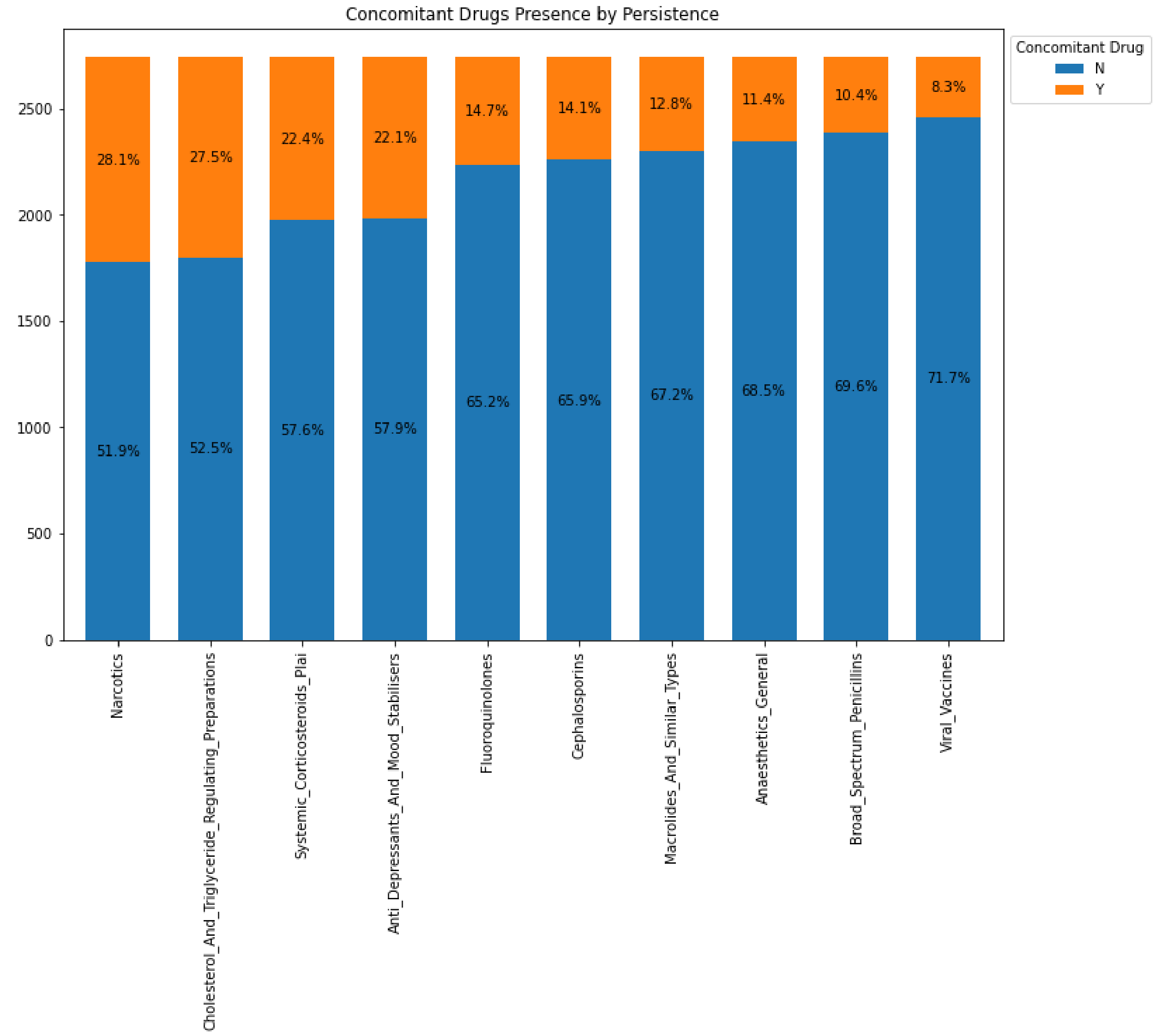
The bar plot presented below shows the **Risk Factors** that are to be considered to affect the Persistency of a patient. We can see that 37.9% of the patients had a Vitamin D defficiency, 14.6% of them smoked, while 11% of the patients are considered with Chronic Malnutrition/Malabsorption.



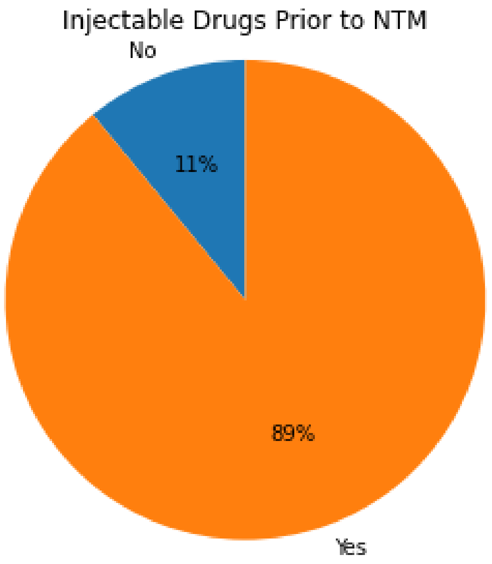
In the bar plot below we have the presence or not of **Comorbidity** the patients of the dataset. Comorbidity is the presence of one or more additional conditions, often co-occurring, with a primary condition. It can indicate either a condition existing simultaneously but independently with another, or a related derivative medical condition. We have 14 different comorbidity factors, with the most prevalent in patients of the dataset being ‘Disorders in Lipoproteins’, in 41.3% of the patients present. Followed by ‘Encounter for Malignant Neoplasms’ and ‘Encounter for Immunization’ in around 35% of the patients present. These factors can be very important as they could increase the rate of NTM infections and so they need to be identified for their importance and treated if possible [**10**].

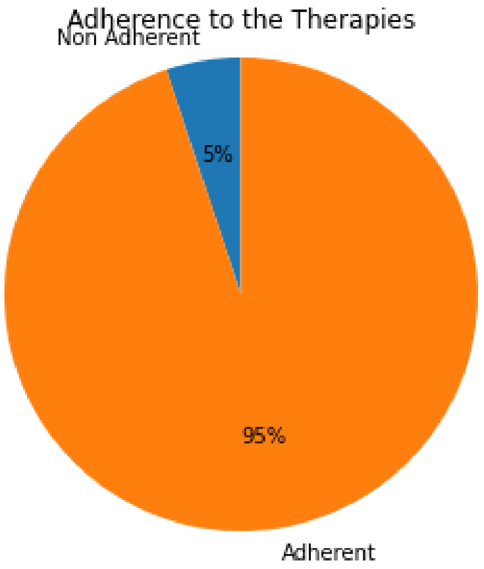


Here we present the **Concomitant Drugs** given to the patients, within 365 days prior to 1st prescription date. Concomitant drugs are two or more drugs used/given *at* or *almost at* the same time. A 28% of the patients had concomitant ‘Narcotics’ closely followed by 27.5% of the patients who had ‘Cholesterol Regulating Preparations’ and in the end we can see the presence of ‘Viral Vaccines’ on only 8.3% of the patients.



In the pie charts below we can see the percentages of patients who had any **Injectable Drug** usage in recent 12 months prior to NTM first prescription and the percentages for **Adherent** patients, i.e. the patient's extent of conformity to the recommendation about day to day treatment. The vast majority of patients, 95% were adherent to the recommendations and the greater part of the patients, 89%, had an Injectable Drug usage in the recent 12 months prior to the beginning of their treatment.





We will continue our Exploratory Data Analysis with Univariate relationships, that is we will evaluate whether or not our predictors are independent from our target variable. If they are, they can then be candidates to be removed from the Feature Selection that will be performed later on. Since we have almost entirely categorical variables in our dataset, 59 of our predictors are binary ('Yes','No'), two are counts ('Count of Risks' and 'Dexa Frequency During Rx') and the rest are also categorical, starting with 2 and all the way to 31 levels ('Ntm Speciality').

That is why we will use Pearson's Chi Squared to evaluate independence, where a contingency table will be computed for each predictor and the target variable and a value following a chi-squared distribution is produced. The significance of that value is also given by the test and according to that significance, whether or not it considered out of the ordinary, we will evaluate the predictor's independence.

Presented below is the table with the variables considered to be independent with our predictor, meaning that the values in their contingency tables appear to be distributed at random.

Variable	**pvalue**
Gender	0.426
Age_Bucket	0.416
Gluco_Record_Prior_Ntm	1.0
Frag_Frac_Prior_Ntm	0.845
Risk_Segment_Prior_Ntm	0.618
Tscore_Bucket_Prior_Ntm	0.314
Risk_Osteogenesis_Imperfecta	1.0
Risk_Untreated_Chronic_Hyperthyroidism	0.716
Risk_Untreated_Early_Menopause	0.716
Risk_Patient_Parent_Fractured_Their_Hip	0.618
Risk_Chronic_Malnutrition_Or_Malabsorption	0.051
Risk_Chronic_Liver_Disease	0.487
Risk_Family_History_Of_Osteoporosis	0.624
Risk_Low_Calcium_Intake	0.933
Risk_Excessive_Thinness	0.086
Risk_Hysterectomy_Oophorectomy	0.408
Risk_Estrogen_Deficiency	0.548
Risk_Recurring_Falls	0.699
Race	0.420
Ethnicity	0.463

We can see two predictors with value 1, so there seems to not be any relationship between them and our target variable. Also ‘Gender’ and ‘Age’ seem to not be related, as well as many of the ‘Risk’ predictor’s categories, along with ‘Ethnicity’ and ‘Race’. So these predictors seem to be candidates to be removed after our Feature Selection process, if they prove to be insignificant there as well.

Treatment of Categorical Variables

As we have already stated there are many categorical variables in our dataset, some binary, others nominal. We will create 'dummy variables' for these predictors, meaning that we will convert the categories into numbers. For each category of a predictor we will create another column, where we will have the value of '1' if that patient belongs to that category, and the value '0' if the patient does not belong to that category. We will do this method of encoding only, called OneHotEncoding, because we do not consider that any predictor's categories have an ordinal relationship. For example, if the categories of a predictor could be thought of as 'having that same distance' from one another, then Category 1 would take value '0', then Category 2 would take value'1' and so on. But this is not an assumption we can make in our dataset and so we consider every category in a predictor as uniquely different from the other category (or categories) in that same predictor.

Although, by doing this type of encoding we will end up with many more variables than our starting dataset. For this reason, we will drop every predictor's last category, meaning we will not create an additional column for that category and so we can avoid the so-called dummy variable trap. We will treat as 'Ordinal' the variables with an 'Unknown' category, which we consider as missing so as to be able to Impute their values with the KNN method of imputation, because it requires numerical values. These variables are 'T Score Bucket Prior Ntm', 'Change in T Score', 'Risk Segment During'. All other variables will be encoding with the One Hot Encoding method.

After the transformation, we will have a train and test dataset of 102 variables each, where only one was considered numerical from the start: 'Dexa Frequency During Rx'. The numerical variable was only standardized using a Robust Scaler, which removes the median and scales the data according to the features Inter Quartile range, so as not to be very much affected by the large values of this feature.

Feature Selection

Having an output variable, 'Persistent' or 'Non Persistent' and almost all our predictor variables as categorical, leaves us with two choices when it comes to selecting our features.

- Chi Square test
- Mutual Information

We saw a Chi Square implementation in the section above, so we will not explain it further here. As far as mututal information is concerned, it comes from the Information Theory world, where Information will quantify how surprising an event is in bits. Low probability events have more information, while higher probability events have less information. We would like to quantify how much information there is in a random variable's probability distribution (Entropy). A skewed distribution has low entropy but a distribution where events have a more balanced probability of happening has a larger entropy.

The entropy of a dataset could be in terms of the probability distribution of observations in the dataset belonging to one class or the other. A dataset with a 50/50 split of samples for the two classes would have a maximum entropy (maximum surprise) of 1 bit, whereas an imbalanced dataset with a split of 10/90 would have a smaller entropy as there would be less surprise for a randomly drawn example from the dataset. In this way, enropy can be used as a calculation of how balanced the distribution of classes happens to be. In our case, for example, where we have 1713 'Non Persistent'(class0) and 1026 'Persistent' (class1) patients out of a total of 2739, we calculate the entropy as:

$$entropy = -\big(class0 * log2(class0) + class1 * log2(class1)\big) = 0.9541$$

which indicates that the values are in between the classes.

We will apply both methods and check all possible numbers of features: from just 1 all the way to 104, the total number of our predictors. Our criteria for choosing the number of predictors will be according to the Accuracy of a Logistic Classification model. As the number of features increases, so does the Accuracy of our classification model. But there is a point where the number of predictors become too many, increasing the complexity of our model and thus decreasing its overall Accuracy. We would like to find an optimal number of features in order to maximize our Accuracy but not have too many predictors. For better understanding we present some results in the table below, where we have various numbers of features and Accuracy of the model for both Chi Square and Mutual Information methods:

Number of Features	Chi Square Accuracy	Mutual Information Accuracy
15	0.8573	0.85348
20	0.86172	0.85847
30	0.86291	0.86380
31	0.86579	0.86281
35	0.86917	0.86850
39	0.86982	0.86623
40	0.87037	0.86798
41	0.87015	0.86249
45	0.86978	0.86520
80	0.86969	0.86867
100	0.87188	0.87133

We can see a similar picture from both methods. Also as the number of features increases, so does the Accuracy, but after a point it decreases again, only to increase toward the end. We should note that out of the 101 (102 – 1, because we cannot compute for 0 features) in 66 cases the Chi Square method Accuracy was greater than Mutual Information, but the differences were not very large and if we round on the 2nd or 3rd decimal point we reach 87% with about 30 predictors or so. Mutual Information was slower than Chi Square, and since Chi Square seems to reach about 87% Accuracy with fewer predictors, we will keep the result from Chi Square method with 31 predictors.

Model Building

Now that we have the number of features we will start building various models and see which one seems to classify our 'Test' set patients best. From now on we will only consider the predictors chosen in the previous step. The methods for our Classification problem will be Logistic Regression, RandomForest, k-nearest neighbors, Gradient Boosting and ExtraTrees.

The Logistic Regression has the capability of modeling our probabilities with a function that produces values between 0 and 1, which in a binary classification problem is very helpful.

The Decision Tree methods are non-parametric supervised learning methods. It uses a tree-like model of decisions and their possible consequences. It is helpful in our situation since it requires little data preparation (such as Encoding) and is able to handle both numerical and categorical variables.

The K-nearest neighbors method classifies a data point according to the classes of its neighbors. The number of neighbors can be tuned to find the optimal solution. It works with numeric values, hence Encoding is need in this method.

The Gradient Boosting Classifier works with an ensemble of weak learners, usually Decision Trees. It is seen as a numerical optimization problem where the goal is to minimize the loss of the model by adding weak learners using a gradient like procedure.

We can check our model's performance by comparing our actual values and the ones we got as predicted values and then count the correct and incorrect predictions. This will be done with the 'test set', the number of observations that were left out of the computation process. We will compare those values with the ones we got from our model and see how accurate we are, or not. We do this with the help of a confusion matrix:

$$\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix}$$

Where:

- TP(True Positives): correctly predicted ones (‘Persistent’),
- TN(True Negatives): correctly predicted zeroes (‘Non Persistent’),
- FN(False Negatives): incorrectly predicted zeroes,
- FP(False Positives): correctly predicted ones

From here we will calculate the Accuracy and Precision based on:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

Accuracy can be seen as the fraction of predictions our model predicted correctly. Precision can be seen as the ability of the Classifier not to label as positive a sample that is negative.

Another metric we can use to determine the diagnostic ability of a binary classifier is the Receiver Operating Characteristic (ROC) curve. It is constructed by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR):

$$\mathrm{TPR}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$$

And

$$\mathrm{FPR}=\mathrm{FP}/(\mathrm{FP}+\mathrm{TN})$$

The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR). Classifiers that give curves closer to the top-left corner indicate a better performance. A random classifier would give points lying along the diagonal (FPR = TPR)(our baseline).

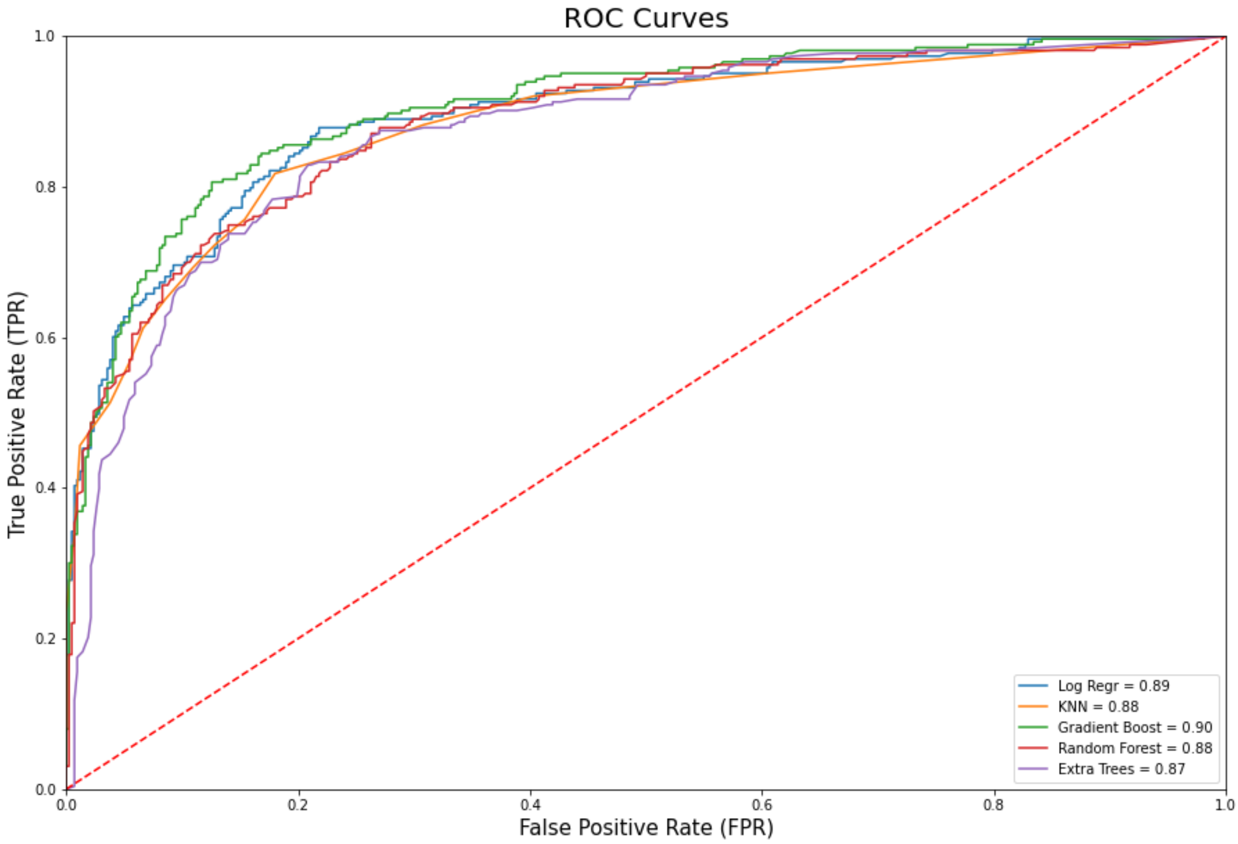
Area Under the Curve (AUC) is used to compare different classifiers, where we summarize the performance of each classifier into a single measurement. AUC measures the entire two-dimensional area underneath the ROC curve. It can be interpreted as the probability that the model ranks a random positive example more highly than a random negative example.

We present a table of our Classifiers and their performance metrics below:

Classification Method	Accuracy	Precision	AUC
Logistic Regression	80.44	80.57	88.55
RandomForest	81.31	81.26	88.44
K Nearest Neighbors	81.31	81.47	88.18
Gradient Boosting	84.10	84.00	90.05
ExtraTrees	80.09	80.08	87.40

From the table we can see that there are not very large differences between the models, when it comes to Accuracy and Precision, GradientBoosting is the best Classification method, while it falls very little behind Logistic Regression in the AUC column. Our recommendation would be the Gradient Boosting Classifier, since we are not interested in coefficients of predictors and their meaning. If that was the case, we would opt for the Logistic Regression Classifier, whose coefficients can be interpreted as logarithms of odds of events.

A plot of the ROC curves of the five models is shown below. No large differences between the models can be observed just by looking at the curves.



References

1. https://www.sciencedirect.com/science/article/pii/S1098301510604950
2. https://curanthealth.com/top-barriers-to-patient-persistence/
3. https://www.pharmexec.com/view/top-barriers-patient-persistence
4. https://faculty.wharton.upenn.edu/wp-content/uploads/2012/04/FS8.Lee-Fader-Hardie.pdf
5. 1976. "Inference and Missing Data." *Biometrika* 63 (3): 581–90
6. https://scikit-learn.org/stable/modules/impute.html#knnimpute
7. https://www.lung.org/lung-health-diseases/lung-disease-lookup/nontuberculous-mycobacteria/diagnosing-and-treating-ntm
8. https://www.nof.org/patients/diagnosis-information/bone-density-examtesting/
9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4470303
10. https://journals.lww.com/md-journal/Fulltext/2019/11080/Incidence,_comorbidities,_and_treatment_patterns.46.aspx