Data Science Project

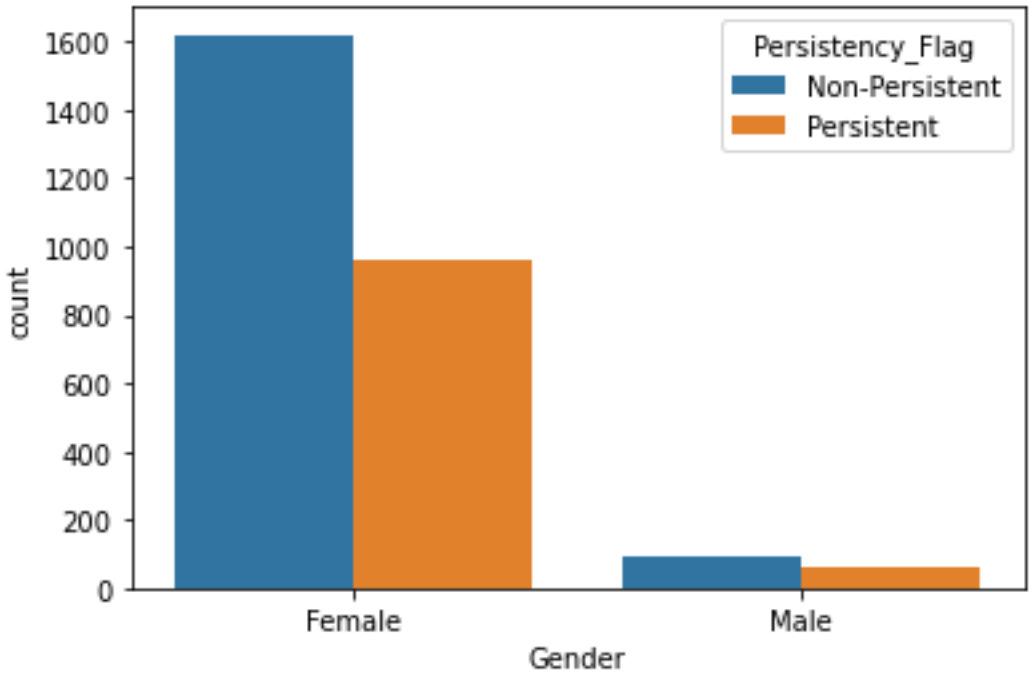
Healthcare - Persistency of a drug

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

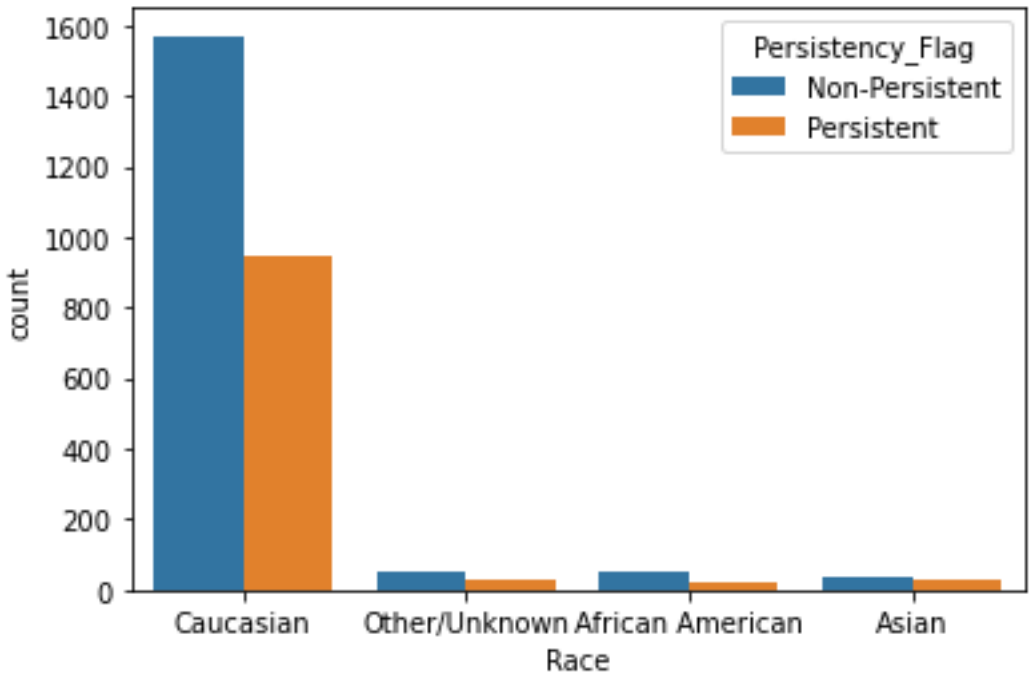
Exploratory Data Analysis

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

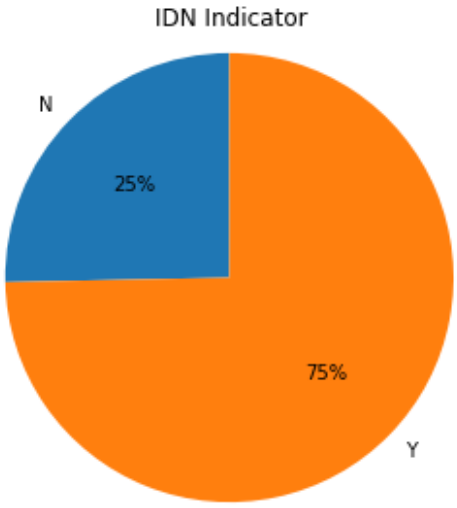
We have below the count plot for the gender of our patients, where 2582 (94.3%) are Female and 157 (5.73%) are Male. For the Female patients, 1620 (62.74%) were non-persistent and 962 (37.36%) were Persistent, while for Male patients, 93 (59.24%) were non-persistent and 64 (40.76%) were Persistent with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



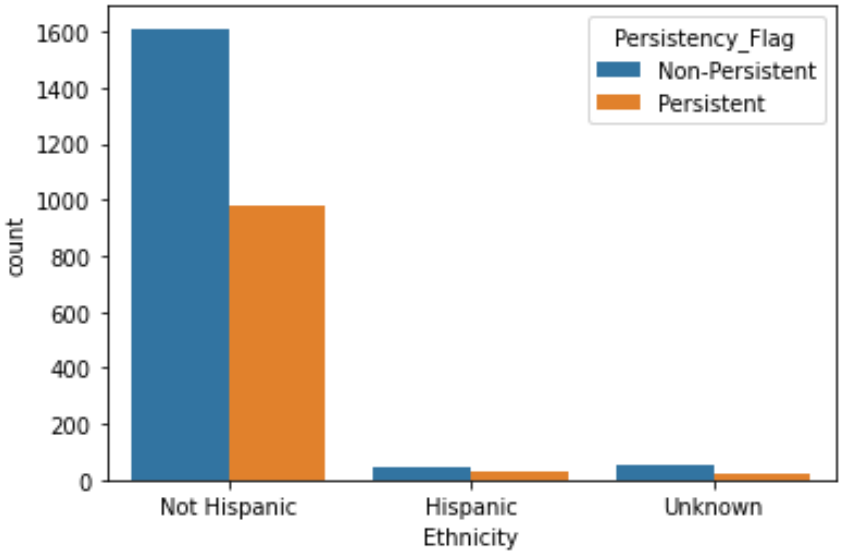
Presented below is the count plot for the Races of our patients and again grouped by their Persistency. We can see that most of our patients were Caucasian (2513, or 91.75%), followed by 'Other/Unknown', then African Americans and then Asians. Again we can see that in all groups we have more non-persistent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



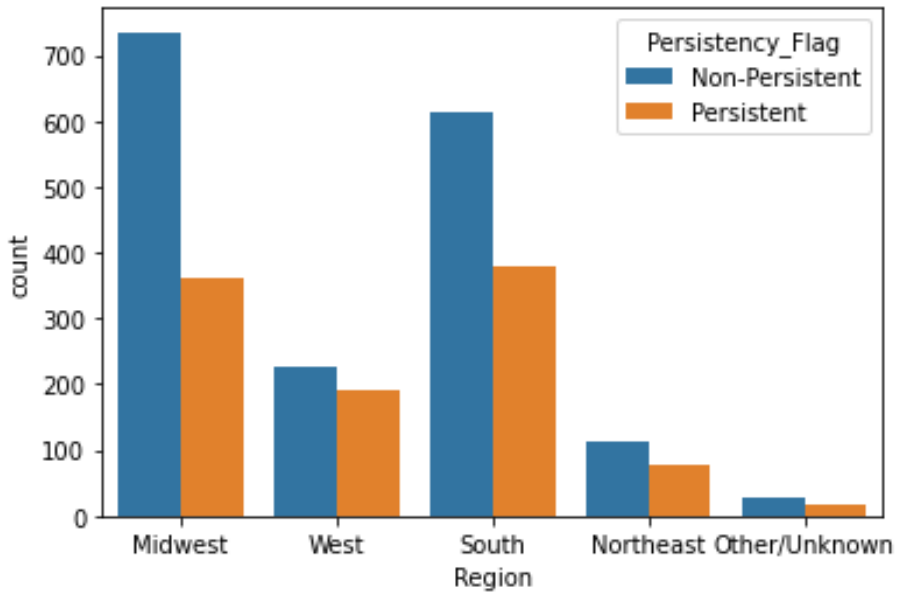
In the pie chart below we have the percentage of patients that were mapped to an **Integrated Delivery Network** (IDN). More than 76% of U.S. hospitals are part of an IDN. Pharma companies research hospitals and care facilities in each of their target IDNs in order to maximize the adoption of their drug/therapy. 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.



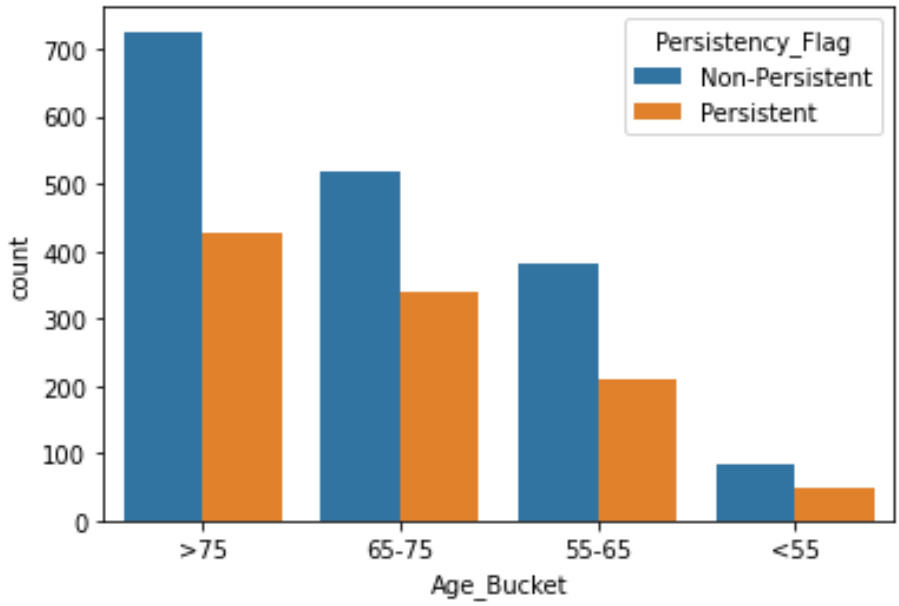
In the figure below, we have the Ethnicities of our patients, with 94.48% being non-Hispanic, 2.7% being Hispanic and 2.8% of 'Unknown' ethnicity. Again we have a picture of more people in each group being non-persistent than persistent with their prescriptions.



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.



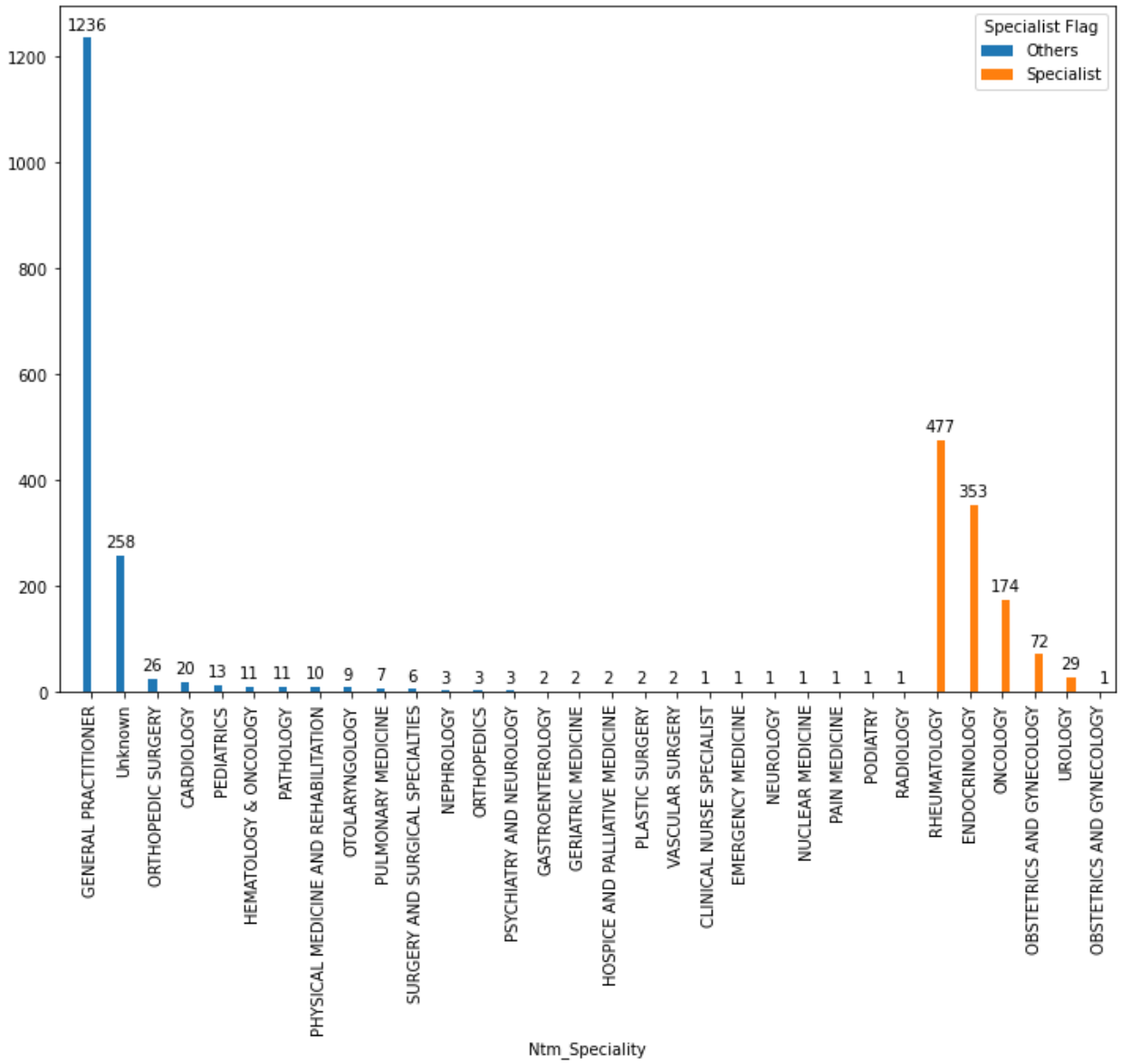
In the image presented below, we can see the age groups of our patients and the drug persistency in each group. Again more patients are non-persistent than they are persistent. We have 42.17% in the category '>75' years old, followed by 31.4% in the category '65-75', then 21.58% in the '55-65' years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

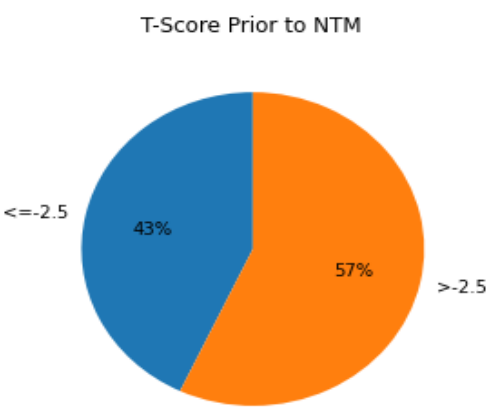


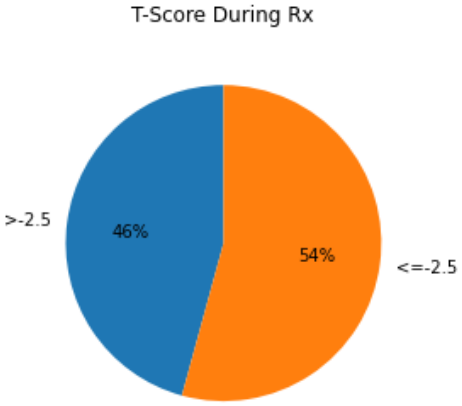
Clinical Factors

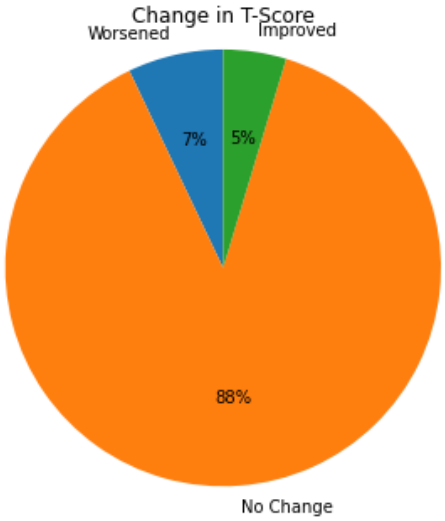
In this bucket of variables we have predictors for the Dexa scans, their results, which are calculated in T-Scores, Fragility Fractures and Glucocorticoid use. We have already seen the Dexa scan variables, so we move on to the **T-Scores** produced by these scans. Every Dexa Scan or a dual-energy X-ray 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.

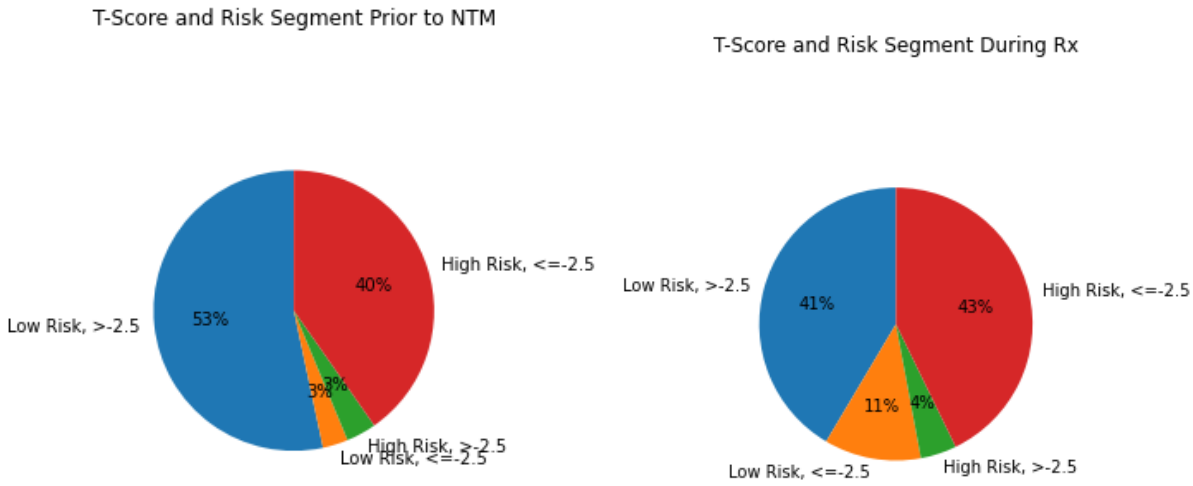






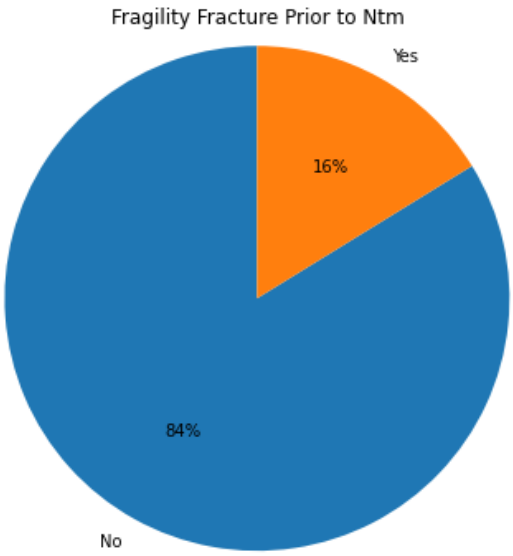
From the figures above we can see that even prior to NTM, 43% patients had a T-Score that indicates low Bone Density. After beginning therapy, we have a large percentage, 54% with low T-Scores. In the last graph, we have the change in the patients' T-Scores before starting any therapy and after receiving 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.

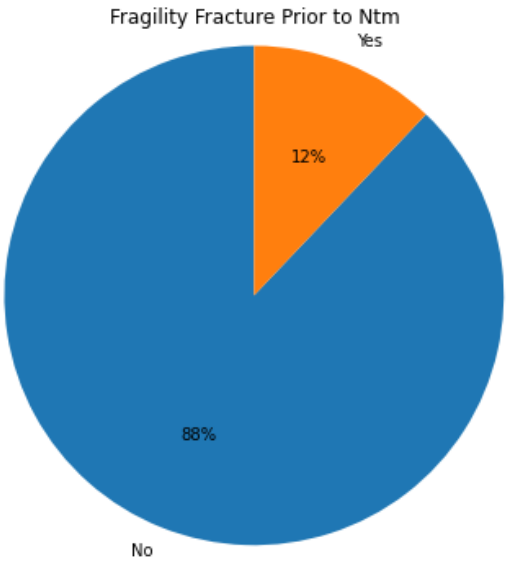
In the figures below, we have plotted the **Risk Segment** variables against the T-Scores obtained for the Dexa scans of the patients.



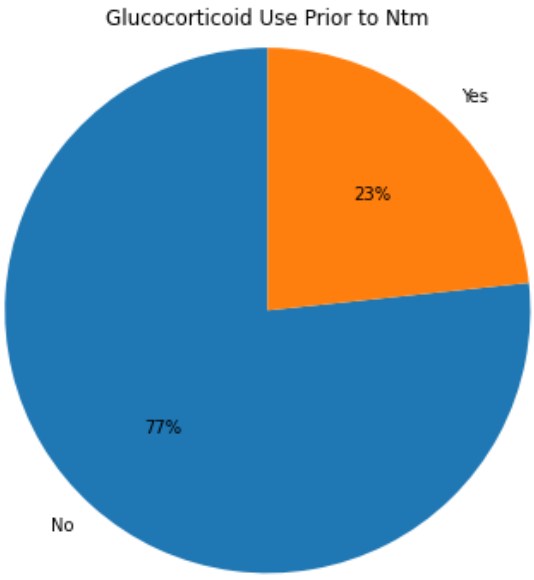
We have combinations of Risk segment and T-Scores in the first two figures. We can see that there is a tendency for these two attributes to correlate, meaning that patients considered 'High Risk' will also have a T-Score value that indicates low bone density and likewise, patients considred 'Low Risk'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

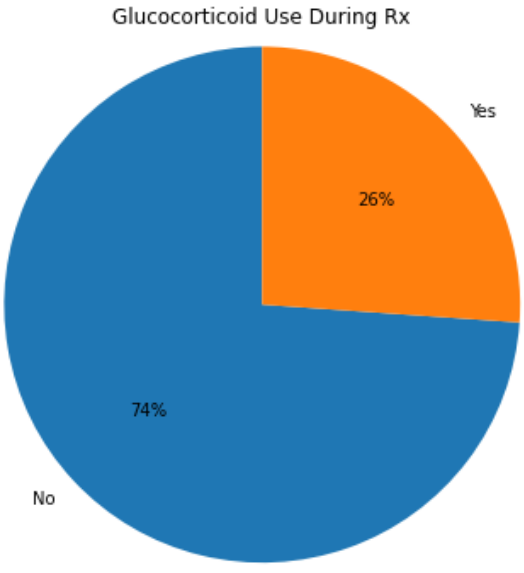
In the two pie charts below, we can see the **Fragility Fracture** predictor, which refers to fractures that result from a fall from a standing height or less. We can see that most our patients, 84% did not exhibit such a Fracture prior to being diagnosed and this percentage actually grew smaller after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





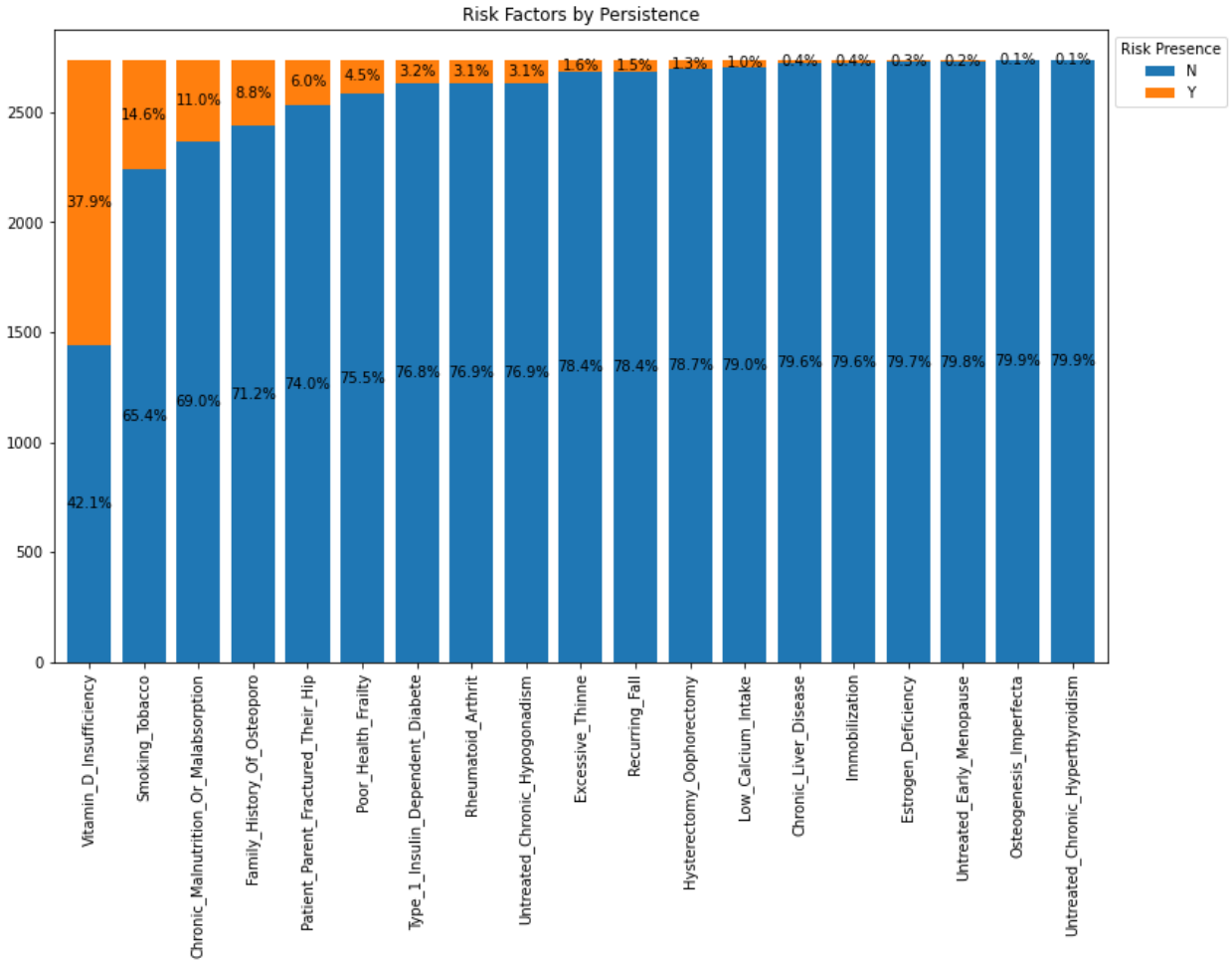
In the figures below we can see **<u>Glucocorticoid Use</u>** from the patients, prior and during their therapies. Glucocorticoids are powerful medicines that fight inflammation and work with the patient's immune system to treat a wide range of health problems. We can see that most patients, 77%, did 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.



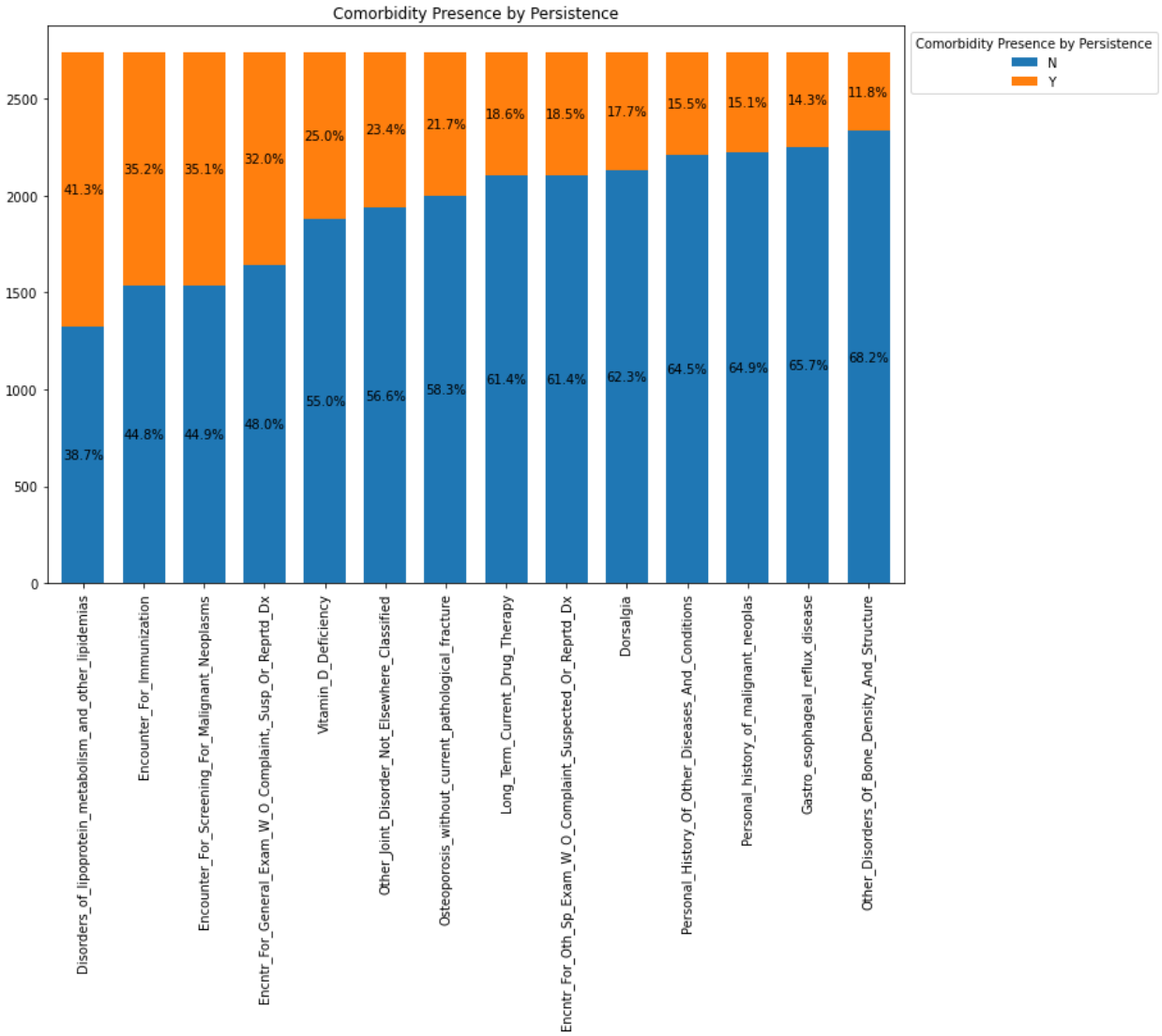


Disease/Treatment Factors

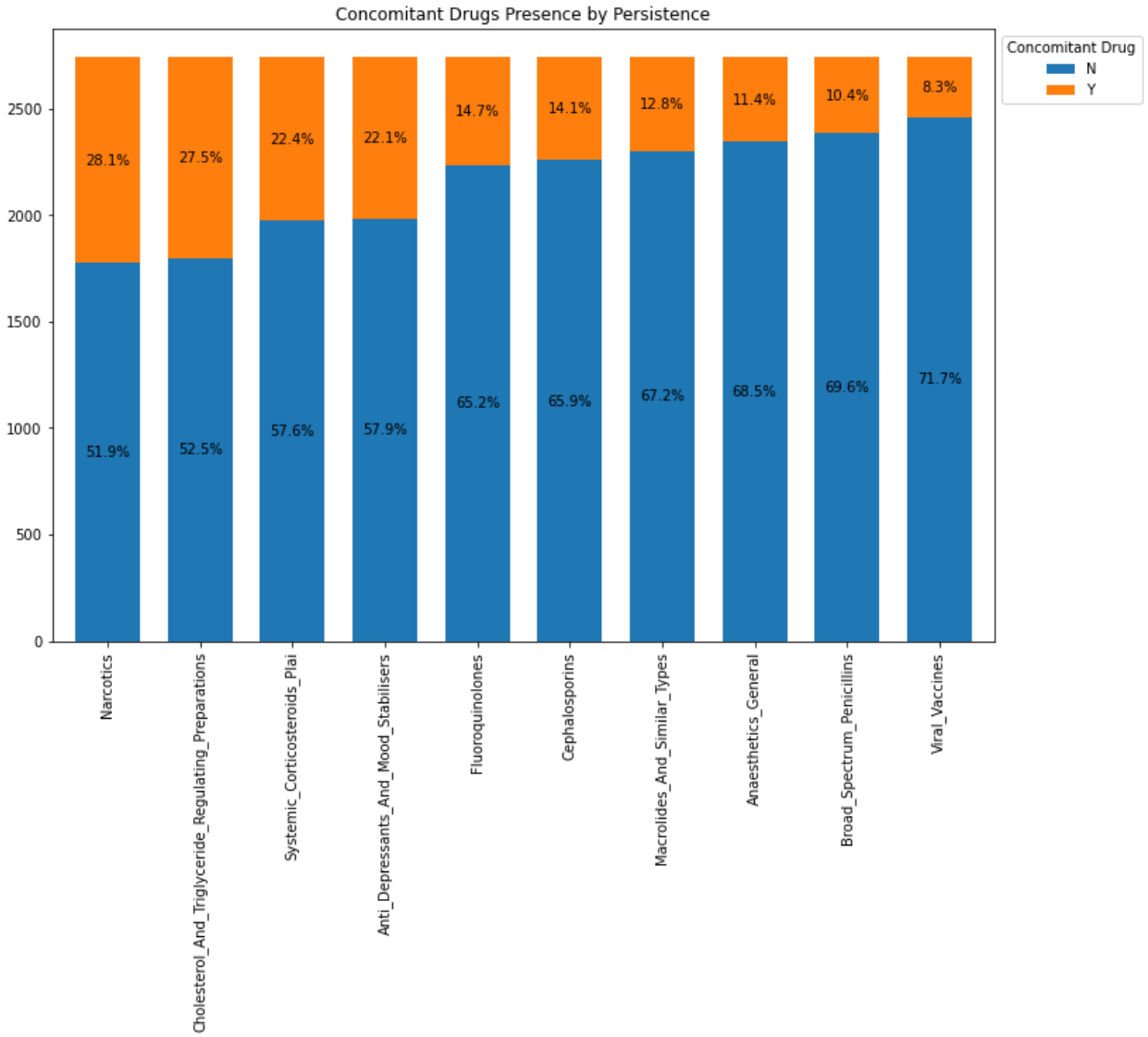
The bar plot presented below shows the **Risk Factors** that are to be considered to affect the Persistency of a patient. We can see that 37.9% of the patients had a Vitamin D defficiency, 14.6% of them smoked, while 11% of the patients are considered with Chronic Malnutrition/Malabsorption.



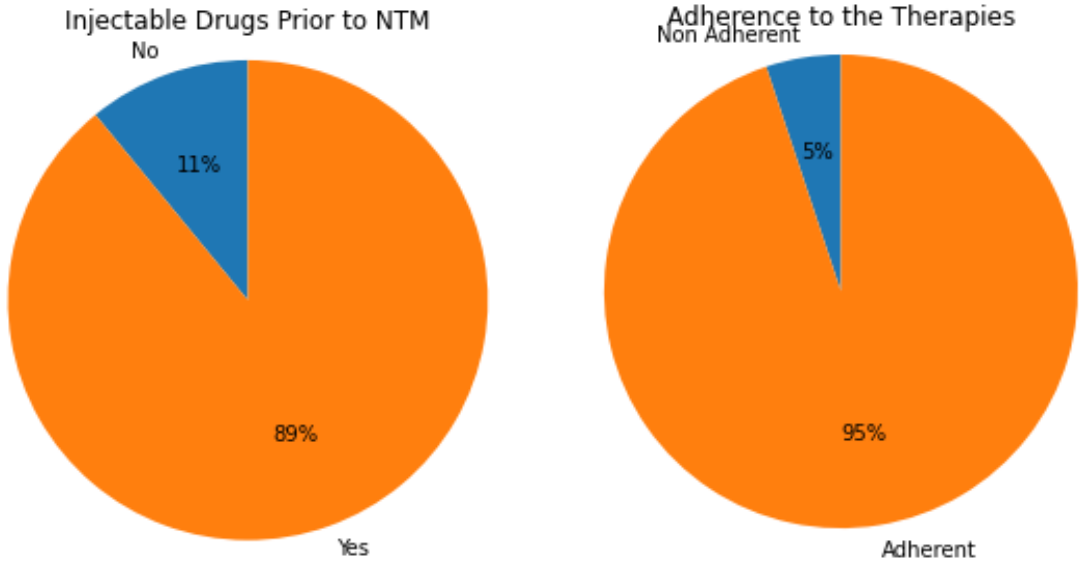
In the bar plot below we have the presence or not of **Comorbidity** the patients of the dataset. Comorbidity is the presence of one or more additional conditions, often co-occurring, with a primary condition. It can indicate either a condition existing simultaneously but independently with another, 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.

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