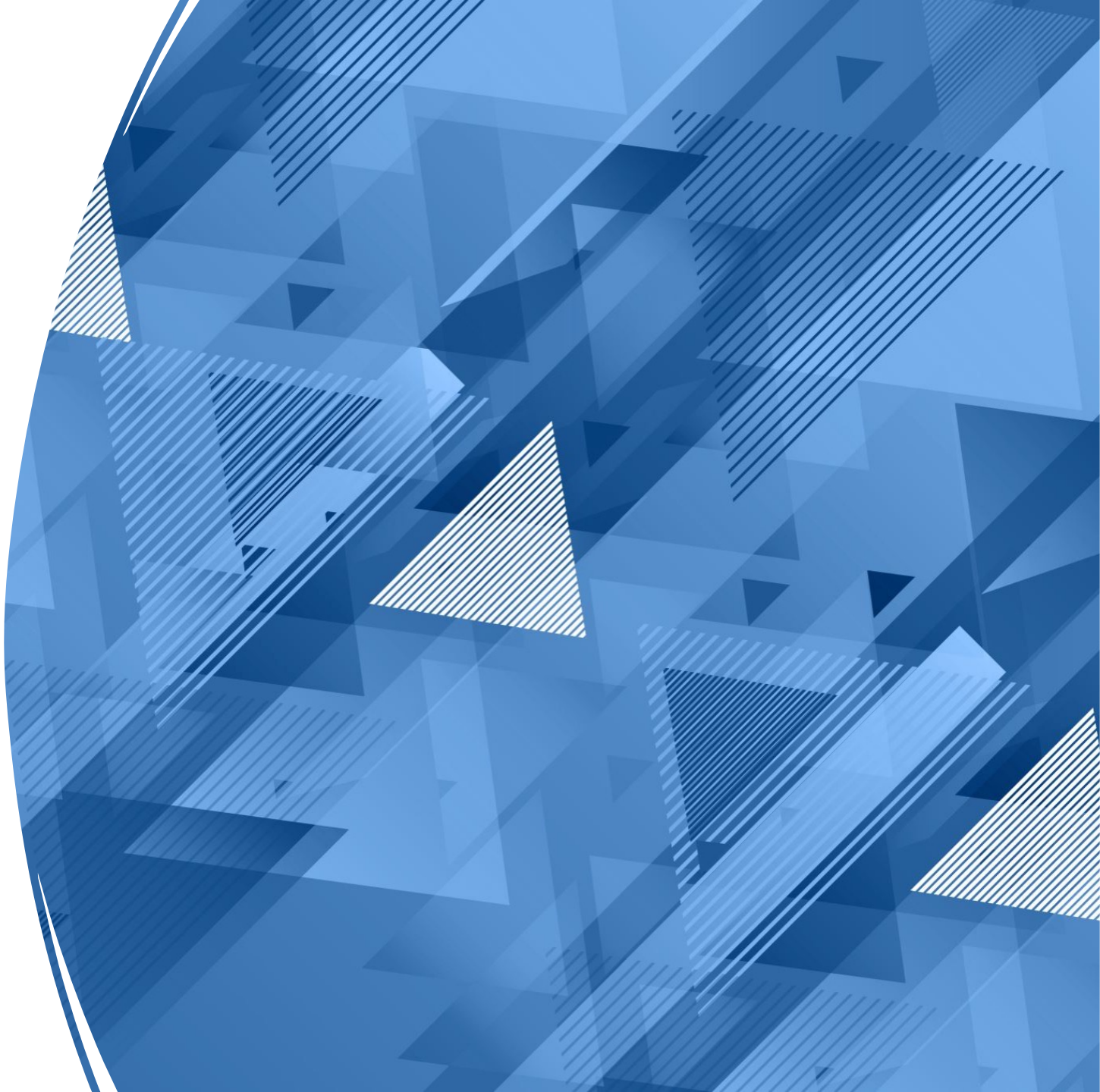
Healthcare Dataset Data Science Specialisation

Exploratory Data Analysis

And Modelling



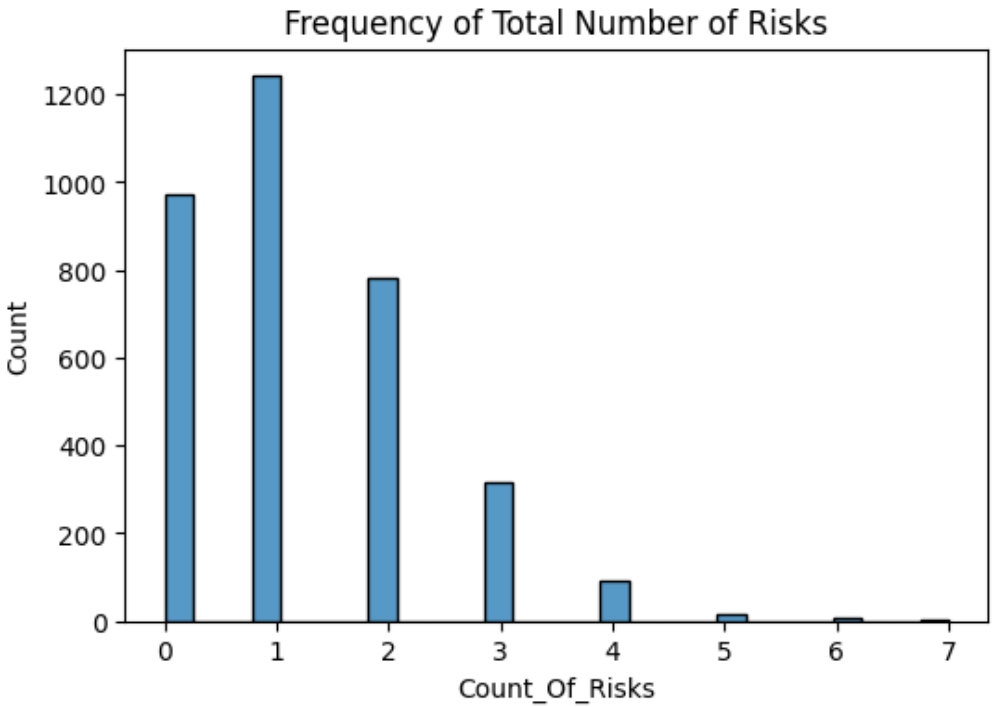
Data Description

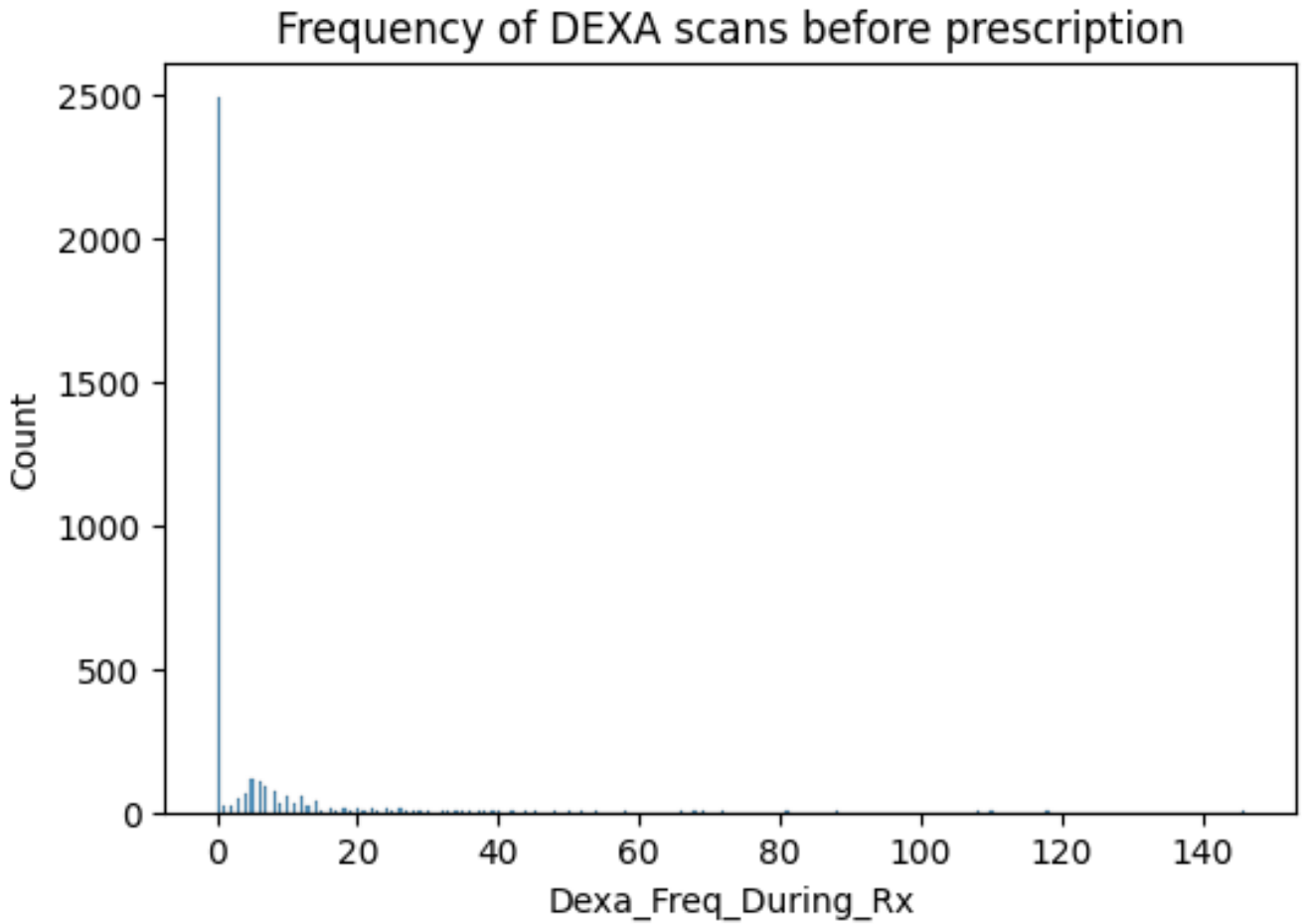
- ABC company (United States) delegated the investigation of persistence of an NTM (non-tuberculous mycobacteria) drug (antibiotic) as a function of various factors
- 3242 participants with data from 67 features including:
  - Patient demographics: Age, gender, ethnicity, etc.
  - Risk factors: Smoking status, obesity, family history, etc.
  - Underlying conditions and comorbidities: Diabetes, hypertension, cardiovascular diseases, etc.
  - Physician information: Specialist or general practitioner, experience level, etc.
- Data mainly categorical (mostly binary variables) and two ordinal variables - frequency of DEXA (bone density) scans and number of pre-existing or ongoing risks for each participant

Data Cleaning

- Data do not contain any NaN/null variables, or any duplicates
- Most categorical variables are imbalanced: some categories of a variable containing hundreds more data points than others
- For such data I removed the variables which presented at least one category that had fewer data points than a specific threshold
- The threshold was determined based on one of the variables used to measure general health: change in risk segment (from high risk to low risk and reverse)
- The risk segment measures the health severity of a person, and in this case also considers the number of serious medical conditions
- The ‘Change in Risk Segment’ variable contains four levels: Unknown (most patients, 2229), No change (1052 patients), Worsened (121) and Improved (22)
- The patients who reported a certain improvement in their risk segment are very few, but sufficient to have their data analysed even without oversampling it or using further alteration methods; thus the number of patients contained in this category represents the threshold for which I eliminated binary variables with categories containing fewer data points than this limit
- Seven such variables were removed as a result
- For ordinal variables I used common statistical elimination methods depending on their skewness: for severely skewed variables such as DEXA scan frequency, I used the flooring and capping method (replacing everything outside of the 25% and 75% quantiles with their corresponding values, give or take the value of the interquartile range, respectively); for the less skewed Count of Risks I replaced everything above the 99% quantile with this value (patients presenting 5, 6 and 7 risks - 23 in total - were newly noted as only presenting 4 risks)

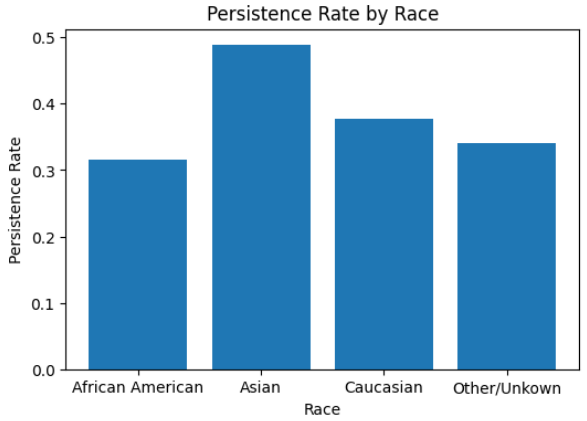
Skewness of Count of Risks and DEXA scan frequency as shown in histograms

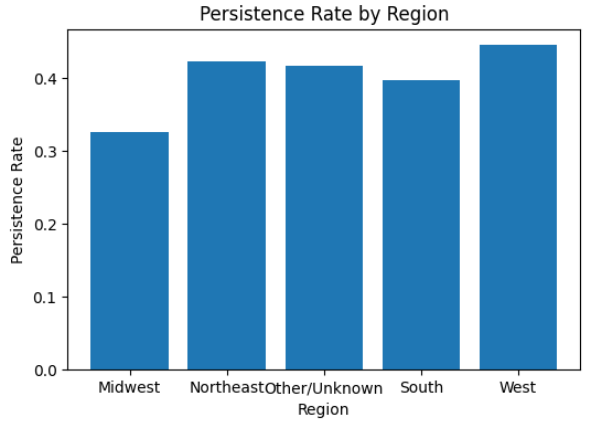


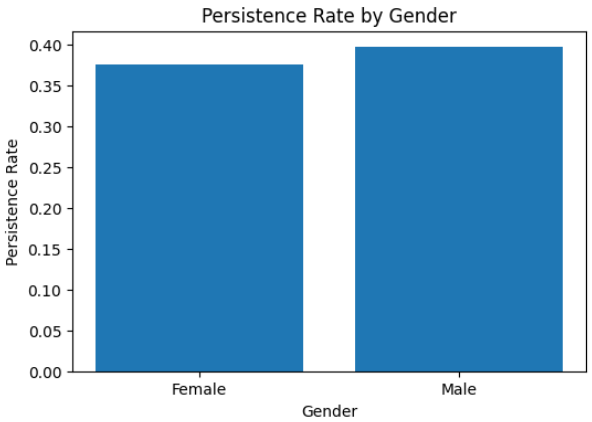


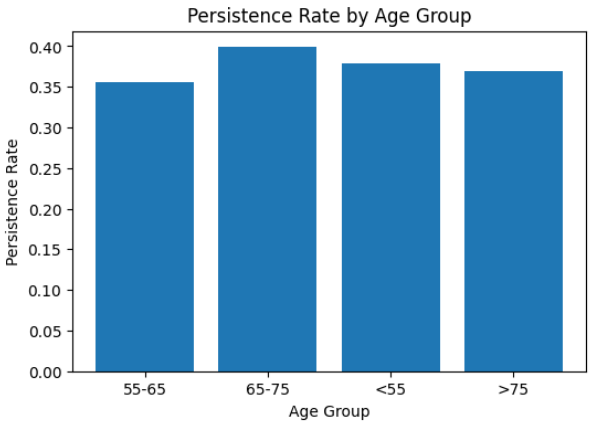
Exploratory Data Analysis

- The challenges in performing exploratory data analysis on the present dataset was the presence of numerous variables, many of them referring to very specific health conditions; knowledge about the interaction of this drug's effects with each of them required high expertise – and it may have been inefficient to investigate further
- Some visualisations are presented for the usual factors: demographics, general information about the prescriber's medical specialisation, and general health information
- I am reporting rates (%) of persistence rather than total numbers of patients showing persistence; to generalise, stats were conducted on this relative persistence; this is because of the number imbalance between the levels of the reported variables. No significant differences were observed at such relative rates – whether this is true of the general population remains unknown, but significant statistical values would have been observed when looking at individual numbers.


Persistence Rate by Race
Persistence Rate
0.0
0.1
0.2
0.3
0.4
0.5
African American
Asian
Caucasian
Other/Unkown
Race


Persistence Rate by Region
Persistence Rate
0.0
0.1
0.2
0.3
0.4
Midwest
Northeast
Other/Unknown
South
West
Region

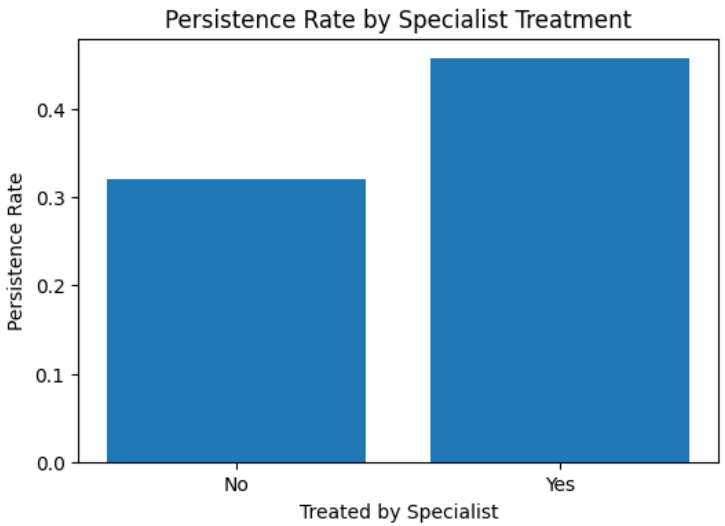

Persistence Rate by Gender
Persistence Rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Female
Male
Gender

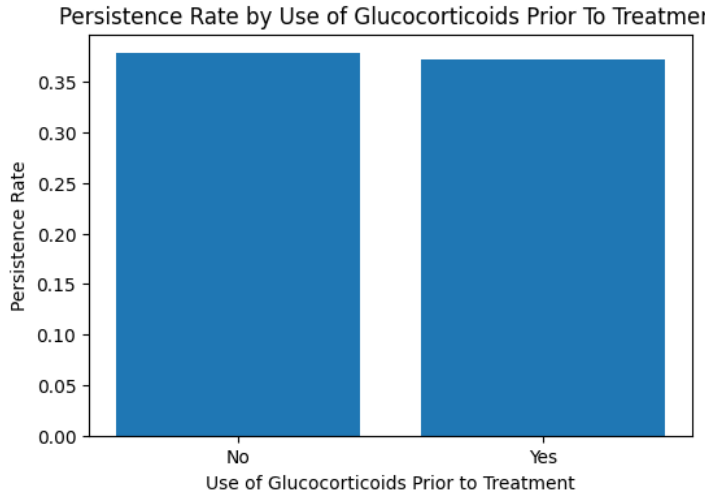

Persistence Rate by Age Group
Persistence Rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
55-65
65-75
<55
>75
Age Group

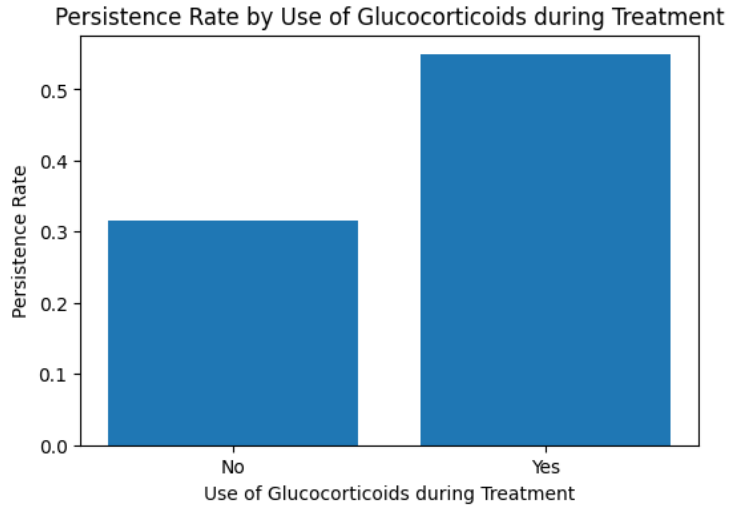
Demographics

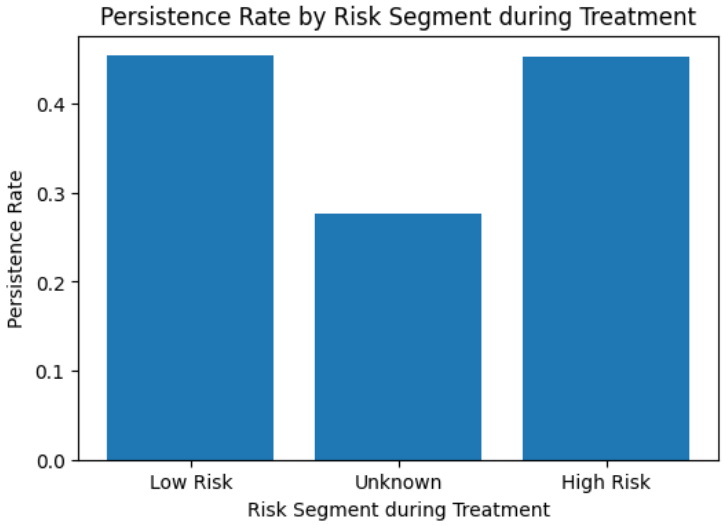
- No noticeable difference in drug persistence between males and females, or between age groups
- Asians show the most drug persistence, followed by the other groups
- As for region, Midwest shows the lowest drug persistence

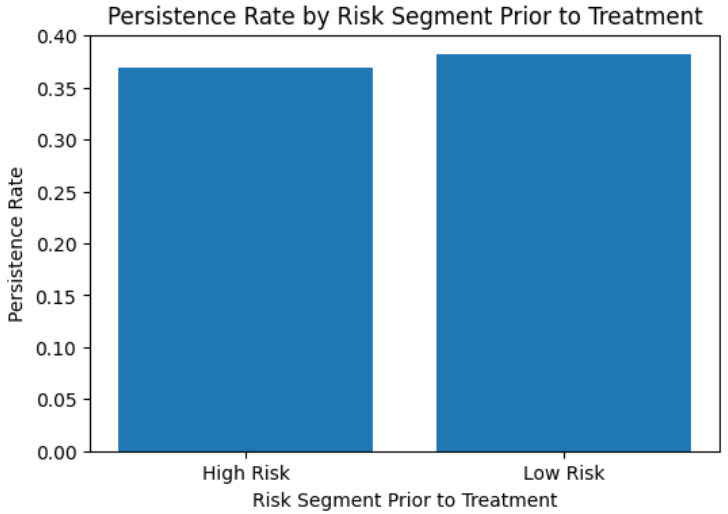
- Persistence rates are higher if antibiotic prescribed by specialist than by GP; in terms of treatment effects, persistence rates higher if antibiotic taken together with glucocorticoids (anti-inflammatory), but not if these were taken before the commencement of the NTM treatment

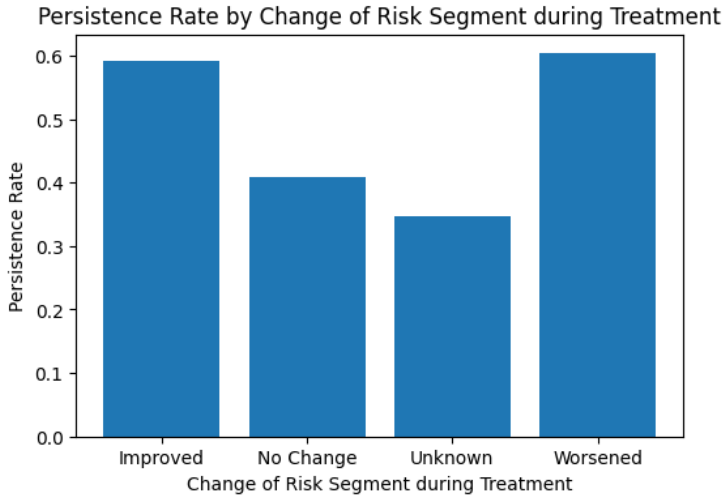






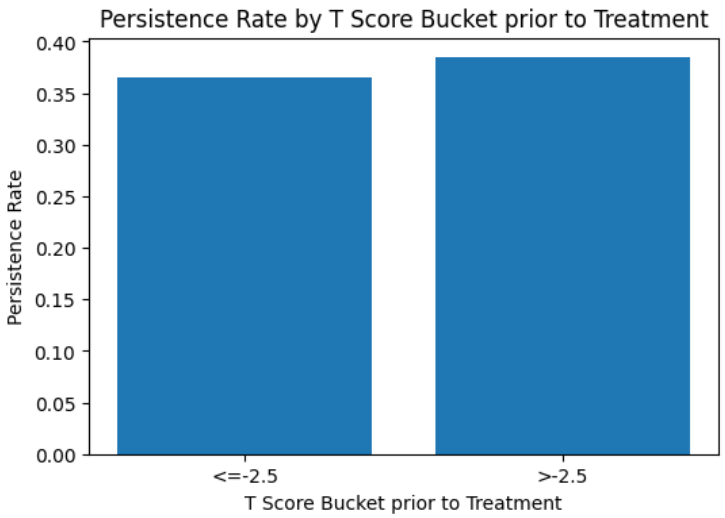


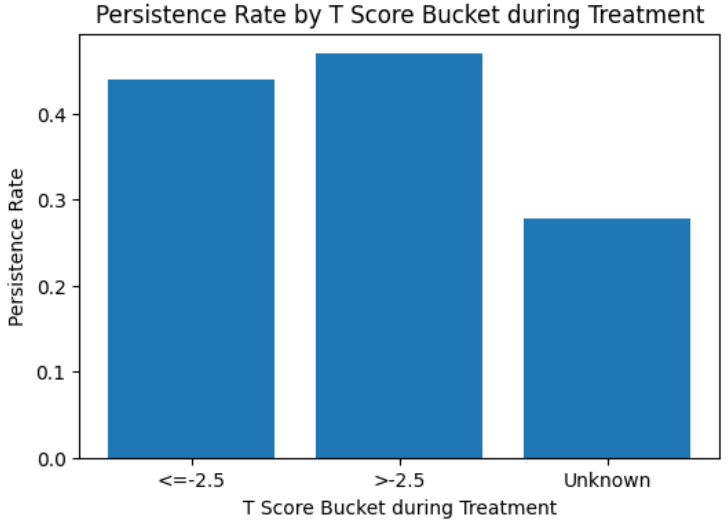


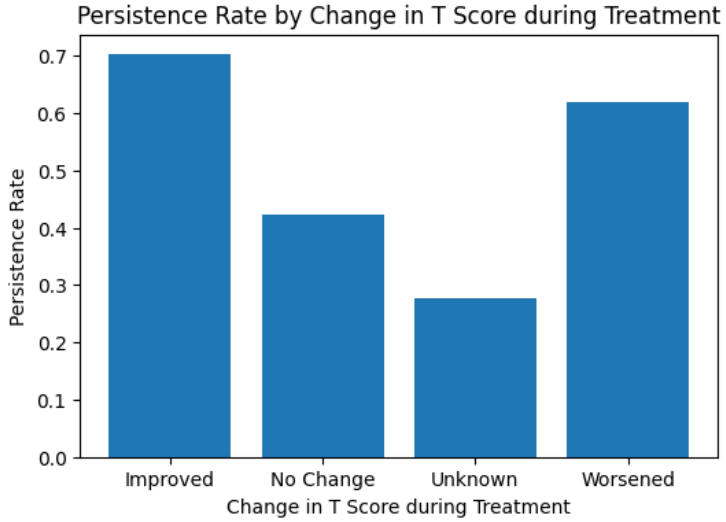


It seems persistence may increase if the risk segment changes – whether this means improving or worsening; but there is no significant difference and the numbers in these categories are quite small

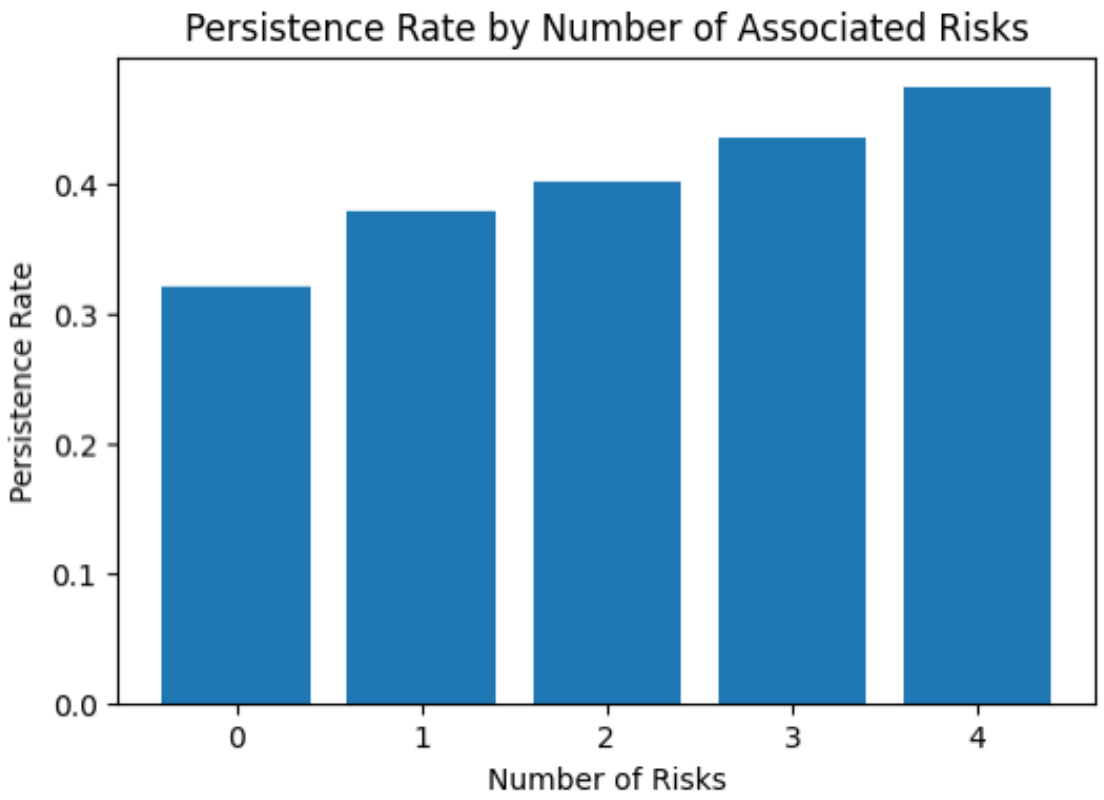
Similar results for change in T Score as for change in risk segment

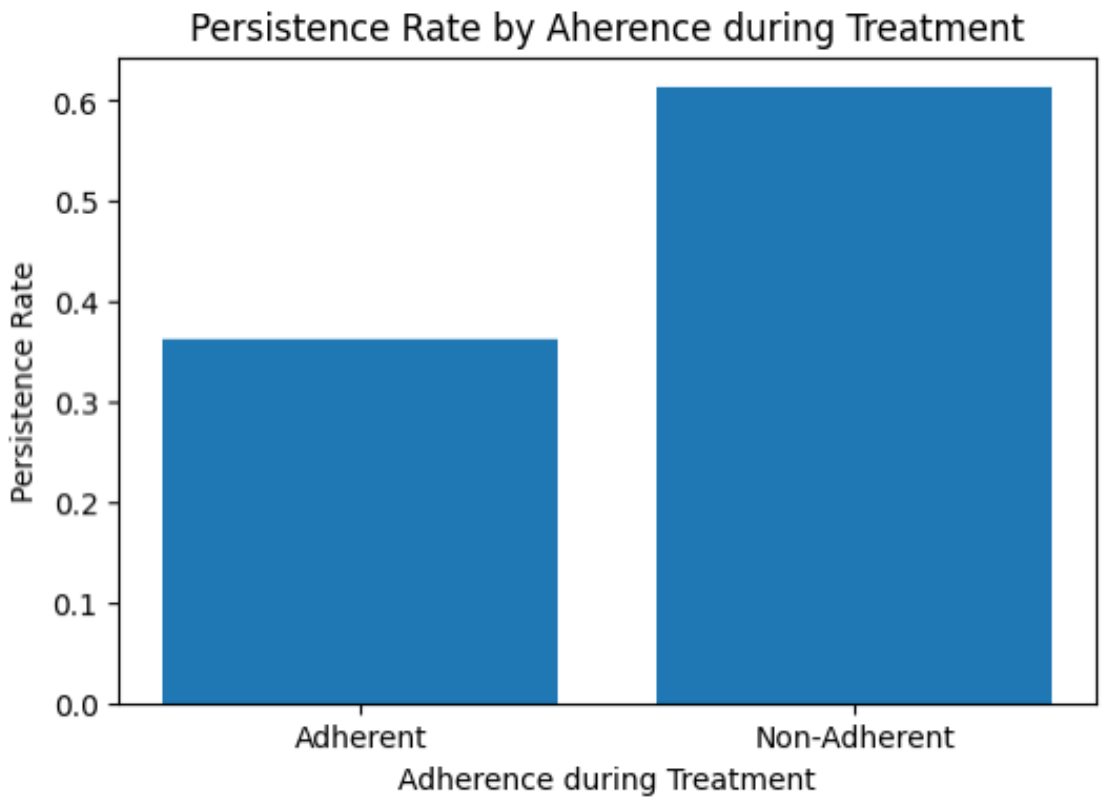






Persistence rates are higher for patients who present risks for higher numbers of conditions, but are lower in patients who show less adherence to treatment (compliance to medical recommendations)



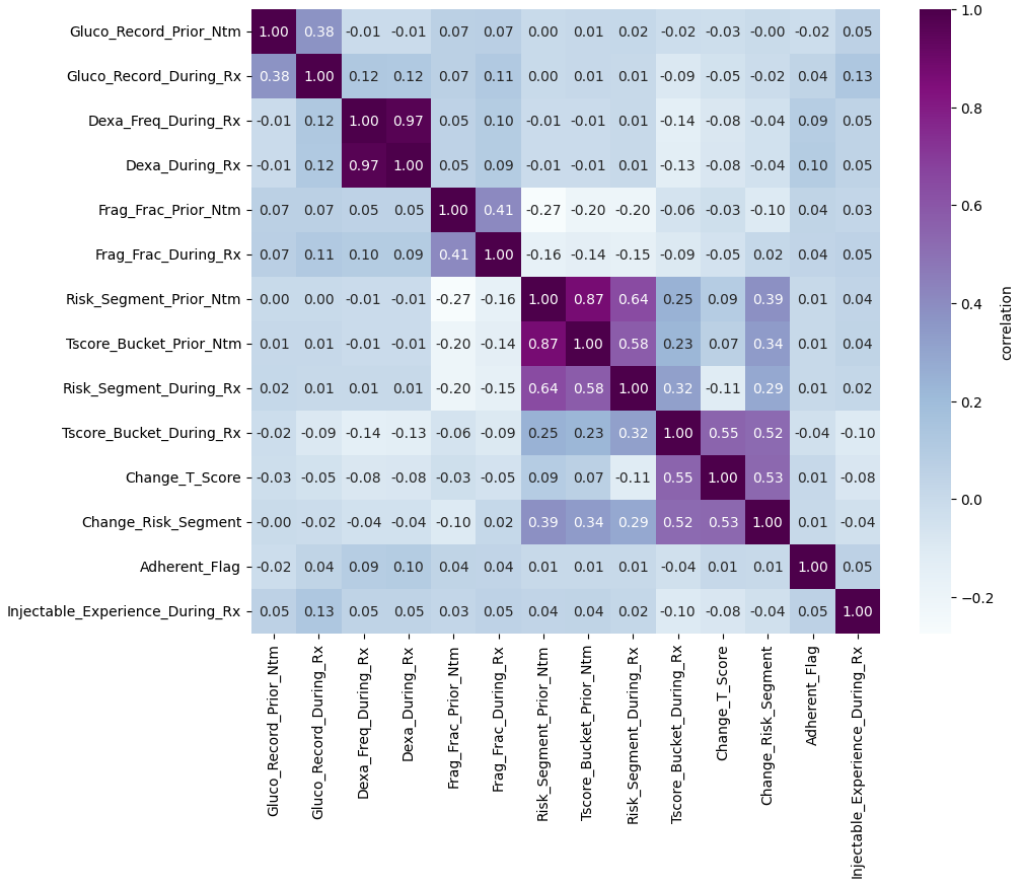


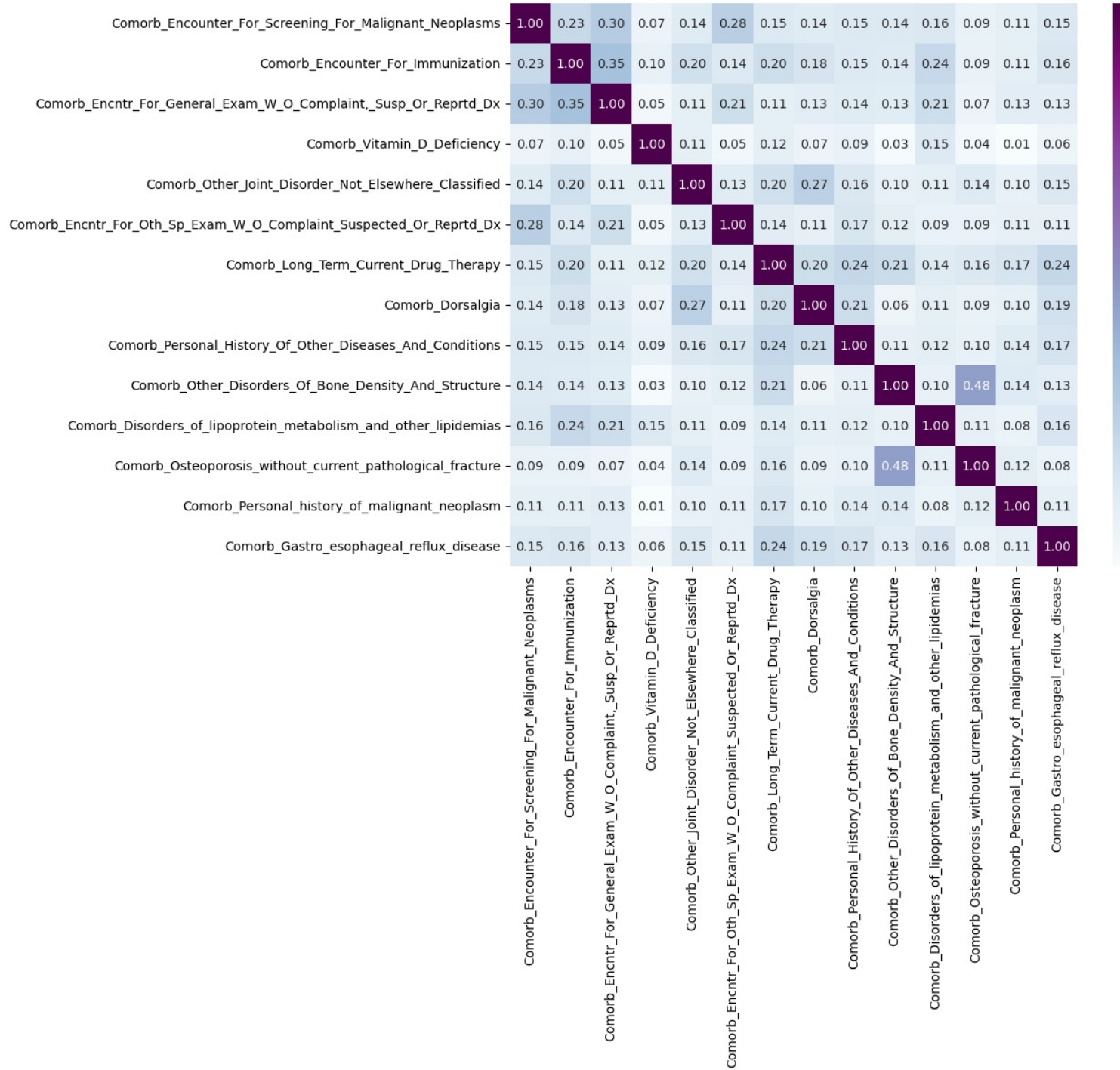
Model recommendations

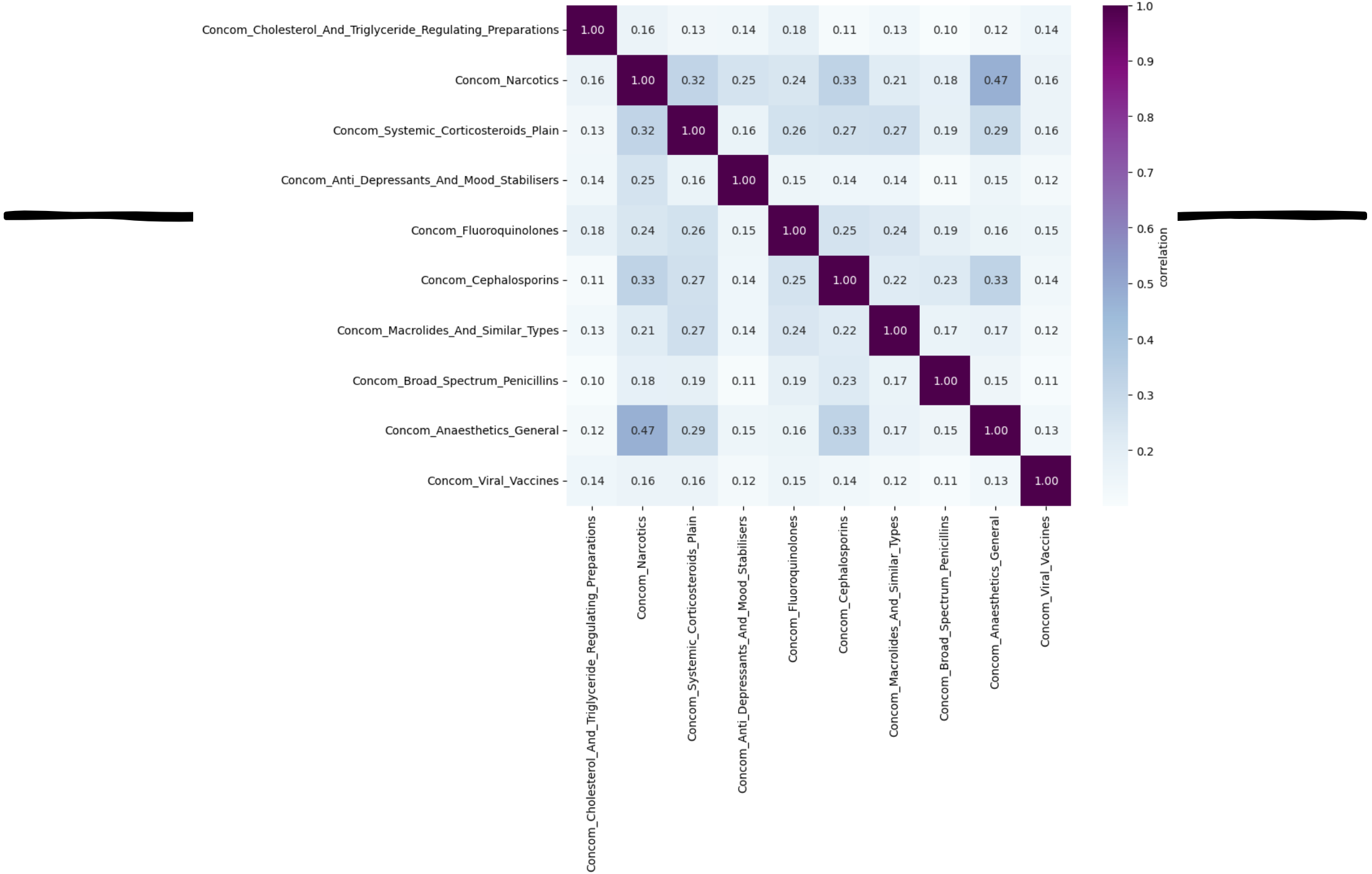
- Dimensionality reduction
- One way to start is by performing correlations on the different categories of variables
- Principal components analysis (PCA) can be used to remove correlated variables
- Recursive feature elimination can be used to remove features one by one and then check best model performance
- Can try oversampling for imbalanced data
- As a general model, logistic regression is recommended for a binary categorical outcome

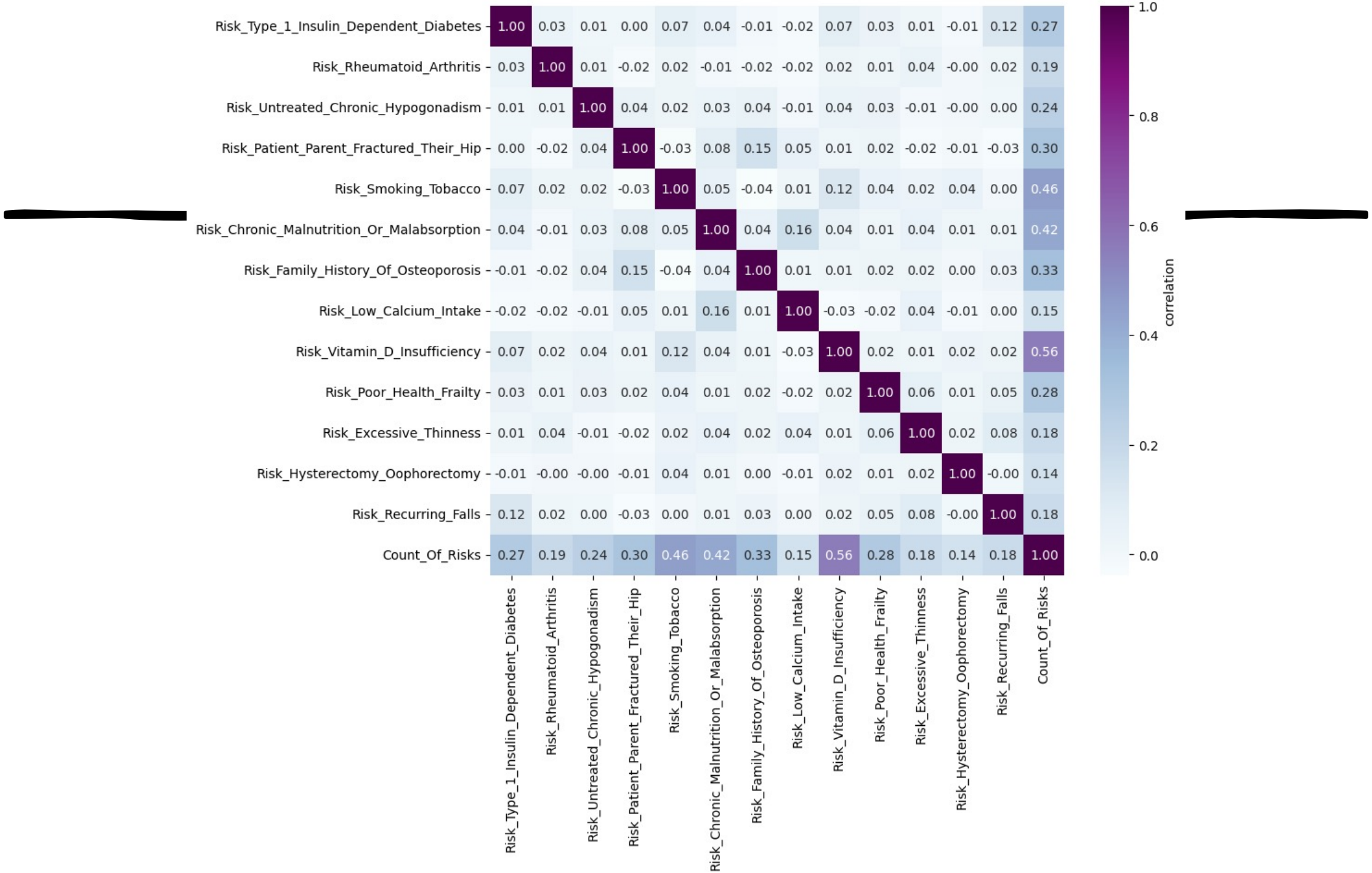
Correlations

- For this I split the variables into their identified categories – mainly for visualisation purposes
- Correlations were observed, at each category:
  - Patient demographics: Not performed.
  - Risk factors: Medium correlations between the total number of risks and being at risk of Vitamin D insufficiency, the risk of smoking tobacco, and being at risk of chronic malnutrition or malabsorption, respectively – possibly indicating that being at risk for these three increases other health risks, but these relationships are outside the current scope
  - Concomitant conditions: Moderate correlation between the concomitant use of narcotics and anaesthetics
  - Comorbidities: Moderate correlation between disorders of bone density and structure and osteoporosis without current pathological fracture
  - Clinical Factors: High correlations between the presence of DEXA scans during treatment and the number of scans; the risk and T scores segments before treatment (but only small during); a relatively high correlation between risks segments before and during treatment; moderate correlation between T score prior to treatment and risk segment during treatment; and medium relationships between the T score, the change in T score and the change in risk segment during treatment.


Gluco_Record_Prior_Ntm 1.00 0.38 -0.01 -0.01 0.07 0.07 0.00 0.01 0.02 -0.02 -0.03 -0.00 -0.02 0.05
Gluco_Record_During_Rx 0.38 1.00 0.12 0.12 0.07 0.11 0.00 0.01 0.01 -0.09 -0.05 -0.02 0.04 0.13
Dexa_Freq_During_Rx -0.01 0.12 1.00 0.97 0.05 0.10 -0.01 -0.01 0.01 -0.14 -0.08 -0.04 0.09 0.05
Dexa_During_Rx -0.01 0.12 0.97 1.00 0.05 0.09 -0.01 -0.01 0.01 -0.13 -0.08 -0.04 0.10 0.05
Frag_Frac_Prior_Ntm 0.07 0.07 0.05 0.05 1.00 0.41 -0.27 -0.20 -0.20 -0.06 -0.03 -0.10 0.04 0.03
Frag_Frac_During_Rx 0.07 0.11 0.10 0.09 0.41 1.00 -0.16 -0.14 -0.15 -0.09 -0.05 0.02 0.04 0.05
Risk_Segment_Prior_Ntm 0.00 0.00 -0.01 -0.01 -0.27 -0.16 1.00 0.87 0.64 0.25 0.09 0.39 0.01 0.04
Tscore_Bucket_Prior_Ntm 0.01 0.01 -0.01 -0.01 -0.20 -0.14 0.87 1.00 0.58 0.23 0.07 0.34 0.01 0.04
Risk_Segment_During_Rx 0.02 0.01 0.01 0.01 -0.20 -0.15 0.64 0.58 1.00 0.32 -0.11 0.29 0.01 0.02
Tscore_Bucket_During_Rx -0.02 -0.09 -0.14 -0.13 -0.06 -0.09 0.25 0.23 0.32 1.00 0.55 0.52 -0.04 -0.10
Change_T_Score -0.03 -0.05 -0.08 -0.08 -0.03 -0.05 0.09 0.07 -0.11 0.55 1.00 0.53 0.01 -0.08
Change_Risk_Segment -0.00 -0.02 -0.04 -0.04 -0.10 0.02 0.39 0.34 0.29 0.52 0.53 1.00 0.01 -0.04
Adherent_Flag -0.02 0.04 0.09 0.10 0.04 0.04 0.01 0.01 0.01 -0.04 0.01 0.01 1.00 0.05
Injectable_Experience_During_Rx 0.05 0.13 0.05 0.05 0.03 0.05 0.04 0.04 0.02 -0.10 -0.08 -0.04 0.05 1.00
Gluco_Record_Prior_Ntm
Gluco_Record_During_Rx
Dexa_Freq_During_Rx
Dexa_During_Rx
Frag_Frac_Prior_Ntm
Frag_Frac_During_Rx
Risk_Segment_Prior_Ntm
Tscore_Bucket_Prior_Ntm
Risk_Segment_During_Rx
Tscore_Bucket_During_Rx
Change_T_Score
Change_Risk_Segment
Adherent_Flag
Injectable_Experience_During_Rx
correlation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Comorb_Encounter_For_Screening_For_Malignant_Neoplasms 1.00 0.23 0.30 0.07 0.14 0.28 0.15 0.14 0.15 0.14 0.16 0.09 0.11 0.15
Comorb_Encounter_For_Immunization 0.23 1.00 0.35 0.10 0.20 0.14 0.20 0.18 0.15 0.14 0.24 0.09 0.11 0.16
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx 0.30 0.35 1.00 0.05 0.11 0.21 0.11 0.13 0.14 0.13 0.21 0.07 0.13 0.13
Comorb_Vitamin_D_Deficiency 0.07 0.10 0.05 1.00 0.11 0.05 0.12 0.07 0.09 0.03 0.15 0.04 0.01 0.06
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified 0.14 0.20 0.11 0.11 1.00 0.13 0.20 0.27 0.16 0.10 0.11 0.14 0.10 0.15
Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx 0.28 0.14 0.21 0.05 0.13 1.00 0.14 0.11 0.17 0.12 0.09 0.09 0.11 0.11
Comorb_Long_Term_Current_Drug_Therapy 0.15 0.20 0.11 0.12 0.20 0.14 1.00 0.20 0.24 0.21 0.14 0.16 0.17 0.24
Comorb_Dorsalgia 0.14 0.18 0.13 0.07 0.27 0.11 0.20 1.00 0.21 0.06 0.11 0.09 0.10 0.19
Comorb_Personal_History_Of_Other_Diseases_And_Conditions 0.15 0.15 0.14 0.09 0.16 0.17 0.24 0.21 1.00 0.11 0.12 0.10 0.14 0.17
Comorb_Other_Disorders_Of_Bone_Density_And_Structure 0.14 0.14 0.13 0.03 0.10 0.12 0.21 0.06 0.11 1.00 0.10 0.48 0.14 0.13
Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias 0.16 0.24 0.21 0.15 0.11 0.09 0.14 0.11 0.12 0.10 1.00 0.11 0.08 0.16
Comorb_Osteoporosis_without_current_pathological_fracture 0.09 0.09 0.07 0.04 0.14 0.09 0.16 0.09 0.10 0.48 0.11 1.00 0.12 0.08
Comorb_Personal_history_of_malignant_neoplasm 0.11 0.11 0.13 0.01 0.10 0.11 0.17 0.10 0.14 0.14 0.08 0.12 1.00 0.11
Comorb_Gastro_esophageal_reflux_disease 0.15 0.16 0.13 0.06 0.15 0.11 0.24 0.19 0.17 0.13 0.16 0.08 0.11 1.00
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms
Comorb_Encounter_For_Immunization
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx
Comorb_Vitamin_D_Deficiency
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified
Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx
Comorb_Long_Term_Current_Drug_Therapy
Comorb_Dorsalgia
Comorb_Personal_History_Of_Other_Diseases_And_Conditions
Comorb_Other_Disorders_Of_Bone_Density_And_Structure
Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias
Comorb_Osteoporosis_without_current_pathological_fracture
Comorb_Personal_history_of_malignant_neoplasm
Comorb_Gastro_esophageal_reflux_disease


Concom_Cholesterol_And_Triglyceride_Regulating_Preparations
Concom_Narcotics
Concom_Systemic_Corticosteroids_Plain
Concom_Anti_Depressants_And_Mood_Stabilisers
Concom_Fluoroquinolones
Concom_Cephalosporins
Concom_Macrolides_And_Similar_Types
Concom_Broad_Spectrum_Penicillins
Concom_Anaesthetics_General
Concom_Viral_Vaccines
1.00 0.16 0.13 0.14 0.18 0.11 0.13 0.10 0.12 0.14
0.16 1.00 0.32 0.25 0.24 0.33 0.21 0.18 0.47 0.16
0.13 0.32 1.00 0.16 0.26 0.27 0.27 0.19 0.29 0.16
0.14 0.25 0.16 1.00 0.15 0.14 0.14 0.11 0.15 0.12
0.18 0.24 0.26 0.15 1.00 0.25 0.24 0.19 0.16 0.15
0.11 0.33 0.27 0.14 0.25 1.00 0.22 0.23 0.33 0.14
0.13 0.21 0.27 0.14 0.24 0.22 1.00 0.17 0.17 0.12
0.10 0.18 0.19 0.11 0.19 0.23 0.17 1.00 0.15 0.11
0.12 0.47 0.29 0.15 0.16 0.33 0.17 0.15 1.00 0.13
0.14 0.16 0.16 0.12 0.15 0.14 0.12 0.11 0.13 1.00
correlation
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

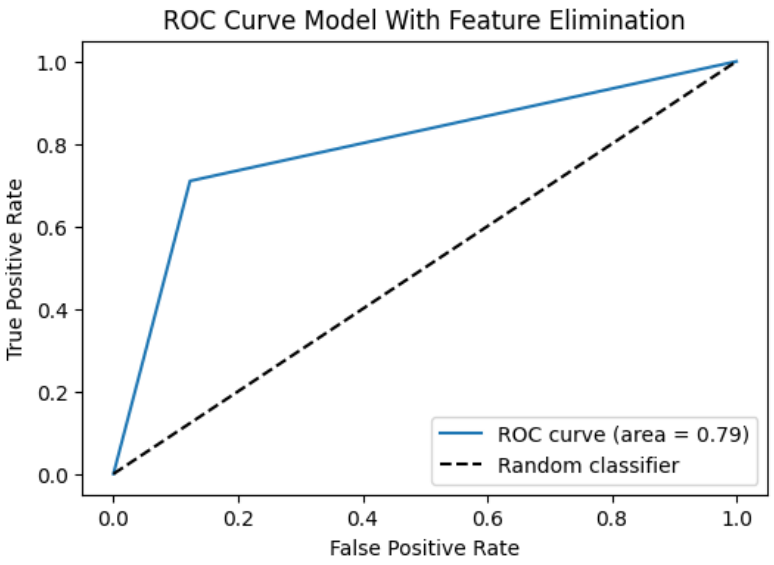

Risk_Type_1_Insulin_Dependent_Diabetes 1.00 0.03 0.01 0.00 0.07 0.04 -0.01 -0.02 0.07 0.03 0.01 -0.01 0.12 0.27
Risk_Rheumatoid_Arthritis 0.03 1.00 0.01 -0.02 0.02 -0.01 -0.02 -0.02 0.02 0.01 0.04 -0.00 0.02 0.19
Risk_Untreated_Chronic_Hypogonadism 0.01 0.01 1.00 0.04 0.02 0.03 0.04 -0.01 0.04 0.03 -0.01 -0.00 0.00 0.24
Risk_Patient_Parent_Fractured_Their_Hip 0.00 -0.02 0.04 1.00 -0.03 0.08 0.15 0.05 0.01 0.02 -0.02 -0.01 -0.03 0.30
Risk_Smoking_Tobacco 0.07 0.02 0.02 -0.03 1.00 0.05 -0.04 0.01 0.12 0.04 0.02 0.04 0.00 0.46
Risk_Chronic_Malnutrition_Or_Malabsorption 0.04 -0.01 0.03 0.08 0.05 1.00 0.04 0.16 0.04 0.01 0.04 0.01 0.01 0.42
Risk_Family_History_Of_Osteoporosis -0.01 -0.02 0.04 0.15 -0.04 0.04 1.00 0.01 0.01 0.02 0.02 0.00 0.03 0.33
Risk_Low_Calcium_Intake -0.02 -0.02 -0.01 0.05 0.01 0.16 0.01 1.00 -0.03 -0.02 0.04 -0.01 0.00 0.15
Risk_Vitamin_D_Insufficiency 0.07 0.02 0.04 0.01 0.12 0.04 0.01 -0.03 1.00 0.02 0.01 0.02 0.02 0.56
Risk_Poor_Health_Frailty 0.03 0.01 0.03 0.02 0.04 0.01 0.02 -0.02 0.02 1.00 0.06 0.01 0.05 0.28
Risk_Excessive_Thinness 0.01 0.04 -0.01 -0.02 0.02 0.04 0.02 0.04 0.01 0.06 1.00 0.02 0.08 0.18
Risk_Hysterectomy_Oophorectomy -0.01 -0.00 -0.00 -0.01 0.04 0.01 0.00 -0.01 0.02 0.01 0.02 1.00 -0.00 0.14
Risk_Recurring_Falls 0.12 0.02 0.00 -0.03 0.00 0.01 0.03 0.00 0.02 0.05 0.08 -0.00 1.00 0.18
Count_Of_Risks 0.27 0.19 0.24 0.30 0.46 0.42 0.33 0.15 0.56 0.28 0.18 0.14 0.18 1.00
Risk_Type_1_Insulin_Dependent_Diabetes
Risk_Rheumatoid_Arthritis
Risk_Untreated_Chronic_Hypogonadism
Risk_Patient_Parent_Fractured_Their_Hip
Risk_Smoking_Tobacco
Risk_Chronic_Malnutrition_Or_Malabsorption
Risk_Family_History_Of_Osteoporosis
Risk_Low_Calcium_Intake
Risk_Vitamin_D_Insufficiency
Risk_Poor_Health_Frailty
Risk_Excessive_Thinness
Risk_Hysterectomy_Oophorectomy
Risk_Recurring_Falls
Count_Of_Risks
correlation
1.0
0.8
0.6
0.4
0.2
0.0

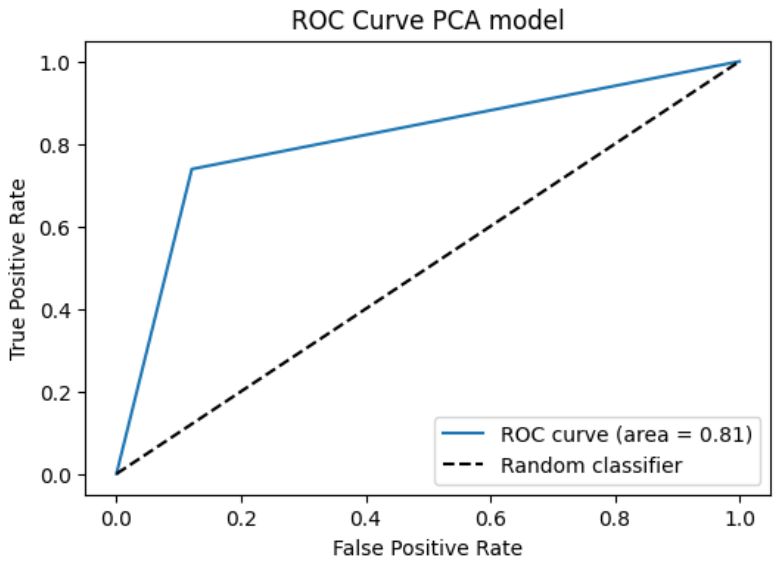
Modelling – logistic regression

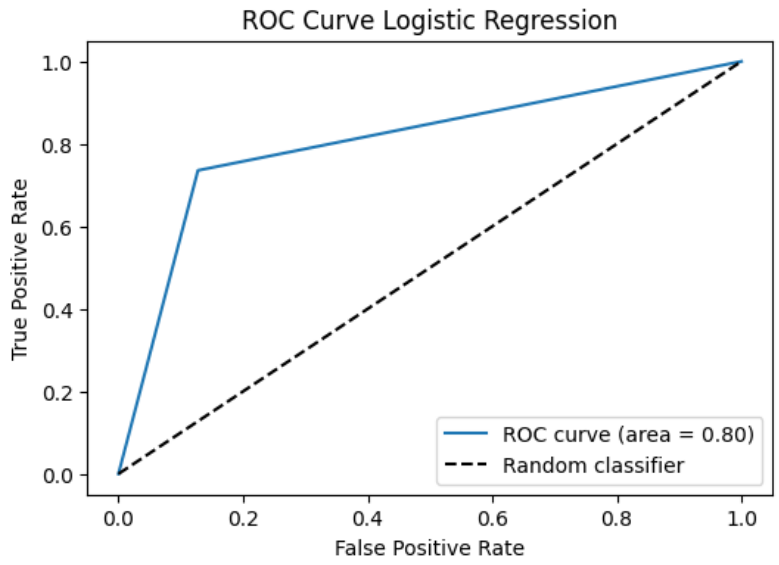
- Logistic regression performed on the cleaned dataset showed between 79 and 84 % accuracy with each run
- So did models on subsequently PCA-ed data (with optimal number of components extracted) or data further reduced using recursive feature elimination
- Oversampling did not lead to a more balanced dataset so was not considered for further modelling

- Model feature importance (normalised data)
- From logistic regression without automated feature elimination

ROC curves for model variations







Conclusion

- Logistic regression was a reasonable chosen model
- All model variations performed equally well, suggesting that data reduction for this particular data set/model was not needed

Future recommendations

- Try same model with even less features (<10)
- Try a different model
- Perform different models for different kinds of data instead of trying to merge them into one (different ones for demographics, risks, etc.)
- Try different ways to identify meaningful and important variables - perhaps by looking at existing scientific results (this could be done either manually or automatically depending on resources)

End