Malignant Melanoma Prediction Using Time-Stamped Hospital Admissions in UK Biobank Participants

David Tang, Malo Dirou, James Parkin, Ross Williamson, Ellie Van Vogt

Health Data Analytics and Machine Learning
Imperial College London

May 2021

Imperial College London

Contents

Introduction

- Risk prediction modelling has previously been applied to various cancer types
- Recent attempts have incorporated multiple approaches including:
  - Phenotype risk scores in prediction of pancreatic cancer
  - Longitudinal data in prediction of pancreatic cancer
  - Polygenic scores in prediction of breast cancers
- Large data stores such as the UK Biobank have facilitated these kinds of analyses, underpin this work, and will continue to support future research

Motivation

- Melanoma is a highly significant cancer
  - Fifth most common cancer in the UK
  - 16,000 cases per year
- Early diagnosis of Melanoma presents a major opportunity
  - 10% of cases are diagnosed at late stage with limited treatment options and one-year median survival of just over 50%
  - However, cases diagnosed at the earliest stage have a 5-year survival of almost 100%
- Despite prior work there is currently no risk prediction model for malignant melanoma in clinical use
  - Kaiser et al, 2020: Systematic review of prior 40 melanoma prediction models over three decades have not included time-stamped comorbidities

Objectives

Primary Aim

To examine whether longitudinal analysis of comorbidities is able to inform time windows for prediction of malignant melanoma.

Secondary Aim

Examination of predictors and comorbidities that are most predictive of malignant melanoma from the most performant machine learning models.

Dataset Descriptions

- UKBiobank Assessment information and measurements (UKB)
  - Demographic and lifestyle information
  - Basic blood measurements
  - Linkage to National Cancer Registry
- Hospital Episode Statistics (HES)
  - Linkage to NHS hospital admissions dating back to 1970s
  - Admissions coding using ICD9 up to 1990s and ICD10 after

Case-Control Selection

Identifying malignant melanoma cases:

- ICD9 code 172, ICD10 code C43
- Cases identified using the NCR variable in the UKB or by hospital admission with codes above
  - 5275 participants identified

Matching Controls:

- Matched controls on a 1:2 ratio
- Matched based on age at recruitment and gender, using a Mahalanobis distance metric
  - This metric takes into account any correlation structure in matching criteria
- Total study population 15825

Risk factors Selection & Data Processing

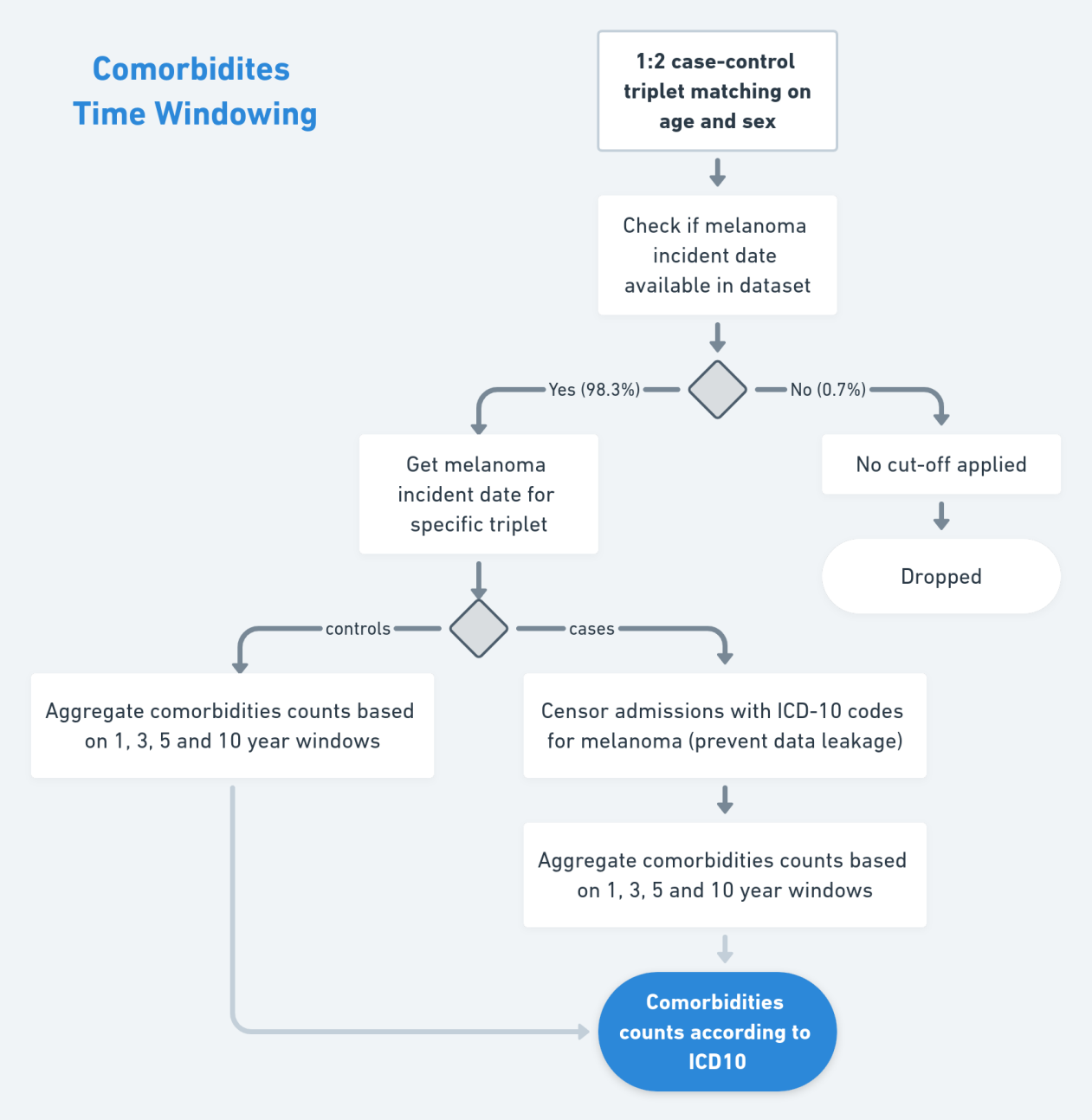
- Literature search to find useful genetic predictors and environmental factors known to be associated with malignant melanomas.
  - Dataset creation:
    - ★ For all 15825 observations:
    - ★ Manual identification of 455 environmental and lifestyle factors in UKB.
    - ★ Appending of comorbidities from HES
  - Data Processing:
    - ★ One-hot encoding of comorbidities
    - ★ Dropped variables whose % missing values for all observations > 40
    - ★ Mean imputation for missing values in numerical variables
    - ★ Most-frequent imputation for missing values in categorical variables when necessary
    - ★ Standardization of numerical variables

Time Windowing Strategy

Identification of diagnosis dates:

- Diagnosis dates identified through earliest entry in HES or UKB diagnosis date, whichever was earliest
  - 156 cases and their paired controls removed due to lack of diagnosis date
  - 1 control excluded due to missclassification
  - 15356 remaining study population
- These dates were then used for each cases matched controls to identify windows of observation

Time Windows



Categorising of comorbidities

Coding:

- Tested keeping ICD10 letters only (A, B, ..), or letters and first number (A0, A1, ..) and letters and 2 numbers (A00, A01, ..)
- A and A0 categorising was too broad, opted to analyse A00 style only.
- for ICD9 codes, kept first 3 characters also

Frequencies of comorbidities:

- exclude variables with less than 0.1% prevalence in study population to exclude rare events
- exclude events relating to pregnancy, other

Results

- Descriptive Statistics
- Separate Logistic Regression Analyses
- Baseline Logistic Regression Model
- Tree-based Models
- Deep Learning

Descriptive Statistics of Study Population

Feature	**Cases**,n = 5275	**Controls**,n = 10550	**Total**,N = 15825
Mean Age (sd)	59 (7.54)	59 (7.54)	59 (7.54)
Sex (%)			
Male	2417 (45.8)	4834 (45.8)	7251 (45.8)
Female	2858 (54.2)	5716 (54.2)	8574 (54.2)
Ethnicity (%)			
White	4947 (93.8)	9346 (88.6)	14293 (90.3)
Other	325 (6.2)	1186 (11.2)	1511 (9.5)
Mean Deprivation Score (sd)	-1.91 (2.69)	-1.33 (3.09)	-1.52 (2.97)
Mean household income pre tax, £ (%)			
$<$ 18k	941 (17.8)	2339 (22.2)	3280 (20.7)
18 – 31k	1182 (22.4)	2380 (22.6)	3562 (22.5)
31 – 52k	1168 (22.1)	2198 (20.8)	3366 (21.3)
52 – 100k	913 (17.3)	1483 (14.1)	2396 (15.1)
$>$ 100k	247 (4.7)	425 (4.0)	672 (4.2)
Skin colour (%)			
Black	0 (0.0)	75 (0.7)	75 (0.7)
Brown	21 (0.4)	266 (2.5)	287 (1.8)
Dark olive	35 (0.7)	167 (1.6)	202 (1.3)
Light olive	577 (10.9)	1900 (18.0)	2477 (15.7)
Fair	3904 (74.0)	7187 (68.1)	11091 (70.1)
Very fair	680 (12.9)	796 (7.5)	1476 (9.3)
Mean childhood sunburns (sd)	2.56 (4.74)	1.42 (3.07)	1.8 (3.71)
Smoking status (%)			
Current	367 (7.0)	1098 (10.4)	1465 (9.3)
Never	3012 (57.1)	5585 (52.9)	8597 (54.3)
Previous	1870 (35.5)	3795 (36.0)	5665 (35.8)
Mean cancer occurrences (sd)	3.61 (2.16)	2.98 (1.45)	3.19 (1.98)
Deaths since enrollment (%)	273 (5.2)	396 (3.8)	669 (4.2)

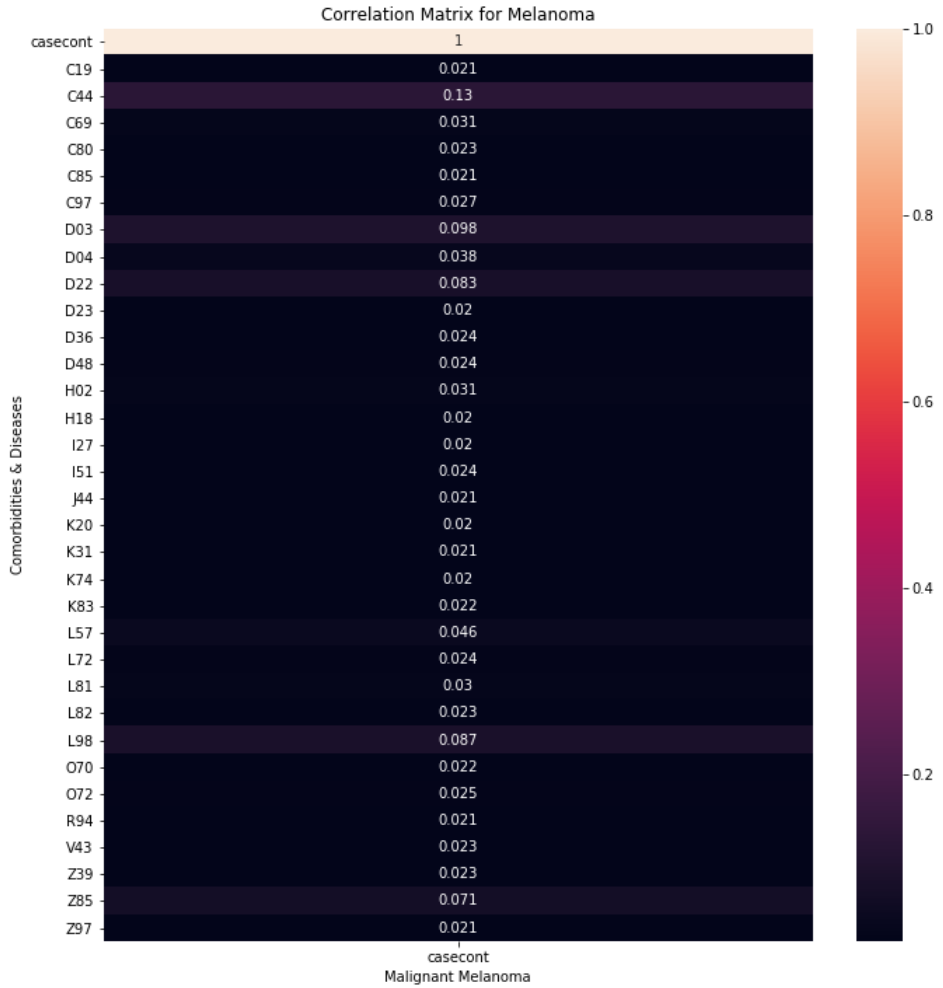
Separate Logistic Regression Analyses

Predictor	Odds Ratio [95% CI]	p-value
Deprivation Score	0.933 [0.922, 0.944]	<0.001
Average Income		
18,000 to 30,999	1	
31,000 to 51,999	1.092 [0.987, 1.207]	0.087
52,000 to 100,000	1.287 [1.151, 1.438]	<0.001
Greater than 100,000	1.218 [1.022, 1.448]	0.027
Ethnicity		
White	1	
Other	0.419 [0.354, 0.493]	<0.001
Smoking status		
Never	1	
Current	0.618 [0.544, 0.7]	<0.001
Previous	0.912 [0.849, 0.98]	0.012
Alcohol Consumption		
Once or twice a week	1	
Daily or almost daily	1.109 [1.006, 1.223]	0.037
Never	0.683 [0.59, 0.788]	<0.001
1-3 times a month	0.984 [0.872, 1.11]	0.794
Special occasions	0.834 [0.738, 0.941]	0.003
3-4 times a week	1.1 [1.001, 1.208]	0.047
Skin Colour		
Fair	1	
Black*	< 0.001 [> 0.001, > 0.001]	0.891
Brown	0.144 [0.09, 0.22]	<0.001
Dark olive	0.383 [0.261, 0.545]	<0.001
Light olive	0.558 [0.504, 0.617]	<0.001
Very fair	1.572 [1.409, 1.755]	<0.001

* 0 vs 75 participants.

Correlation Matrix of Comorbidities

- For due diligence, ensuring no significant and unexpected correlations between malignant melanoma ICD-10 code and other ICD-10 codes exist.



Summary of Models

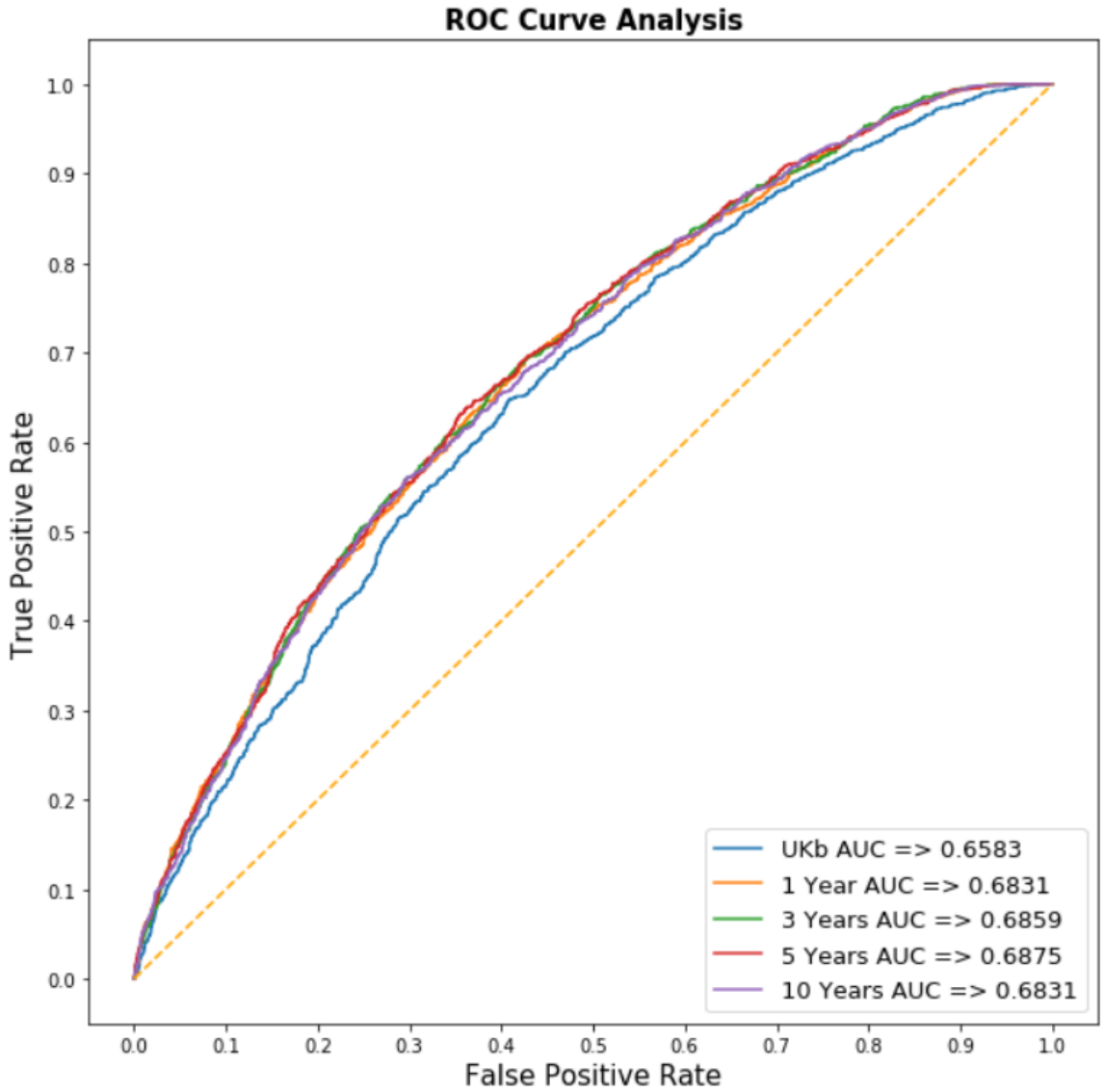
Model	Time-window	AUC
Logistic Regression (UKB)	NA	0.528
	1 year	0.528
	3 years	0.528
	5 years	0.528
	10 years	0.528
Random Forest (UKB)	NA	0.636
	1 year	0.652
	3 years	0.646
	5 years	0.636
	10 years	0.653
XGBoost (UKB)	NA	0.658
	1 year	0.683
	3 years	0.686
	5 years	0.688
	10 years	0.683
XGBoost (PCA)	NA	0.619
	1 year	0.643
	3 years	0.642
	5 years	0.644
	10 years	0.636
Neural Network (UKB)	NA	0.747
	1 year	0.766
	3 years	0.767
	5 years	0.768
	10 years	0.771

NA = no time-windowed comorbidities used

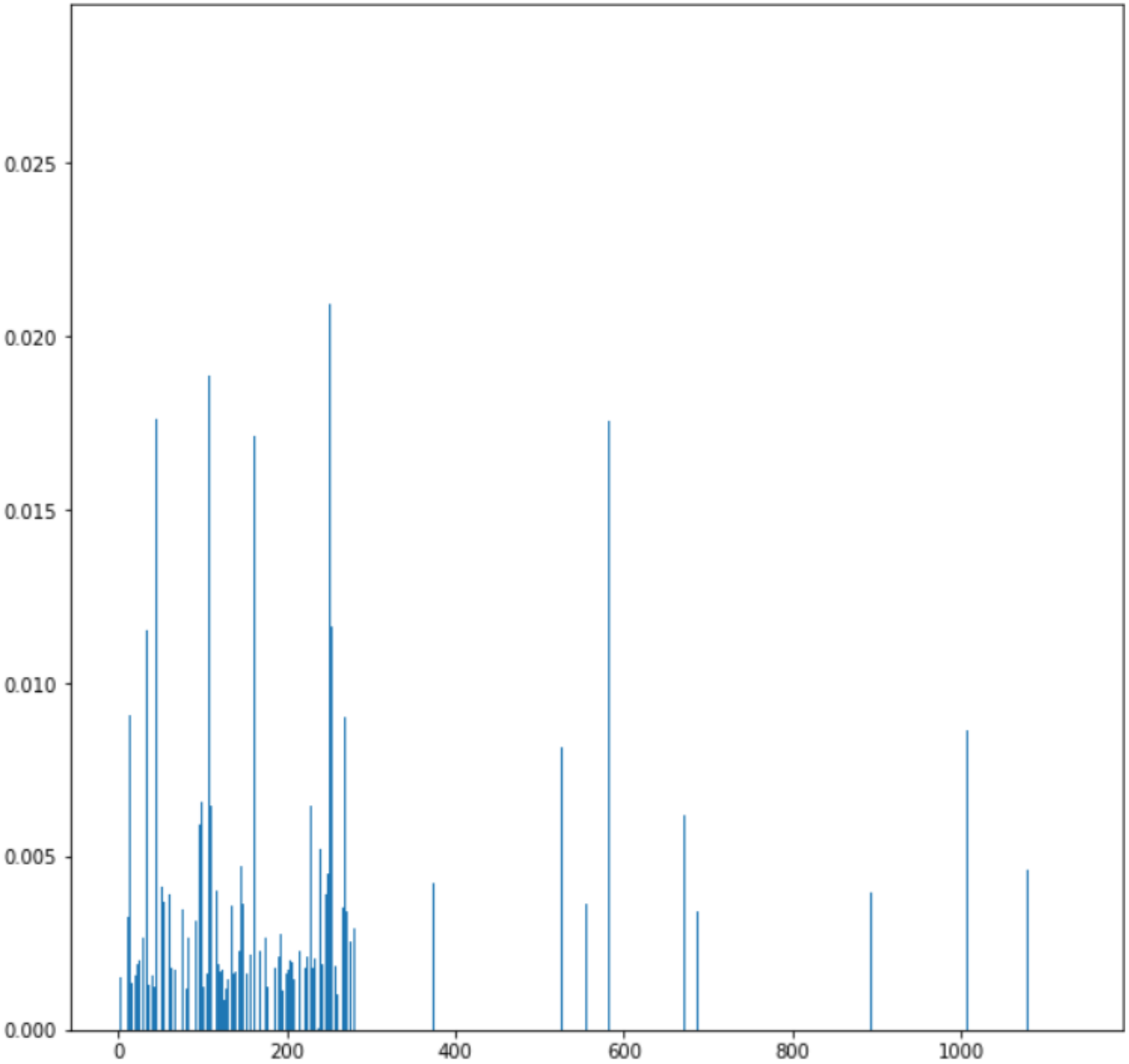
Tree-based methods

- Tree-based models were used in an attempt to improve predictive performance
- Tuning method
  - Manual tuning of max depth monitoring overfitting
  - Brute force gridsearchCV for remaining parameters
- They outperformed baseline logistic regression
- Using the PCA features (XGBoost only) we lost predictive performance
- Initial improvement in 1 year time window
  - Further time windows did not improve performance

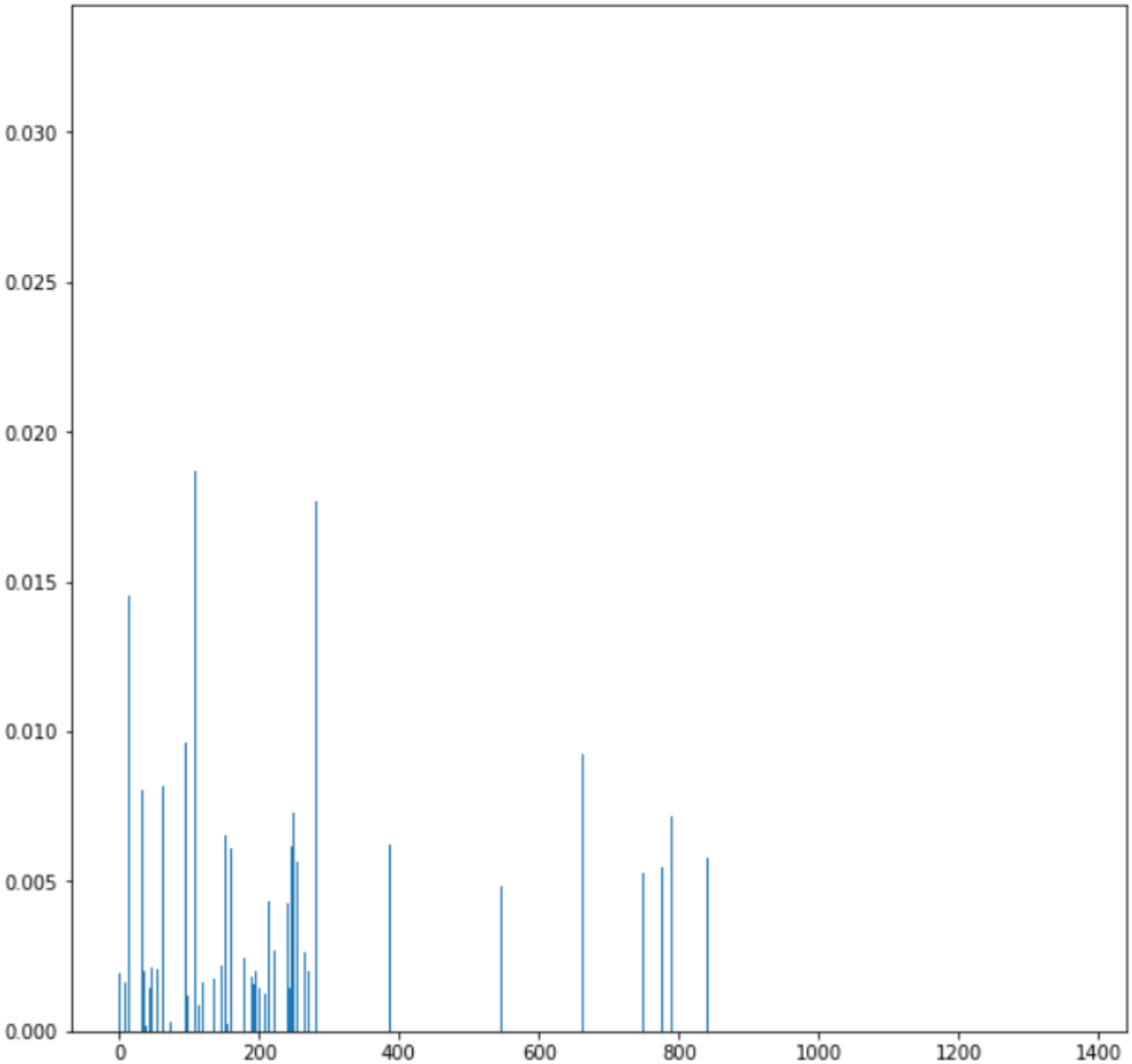
XGBoost


ROC Curve Analysis
True Positive Rate
False Positive Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
UKb AUC => 0.6583
1 Year AUC => 0.6831
3 Years AUC => 0.6859
5 Years AUC => 0.6875
10 Years AUC => 0.6831

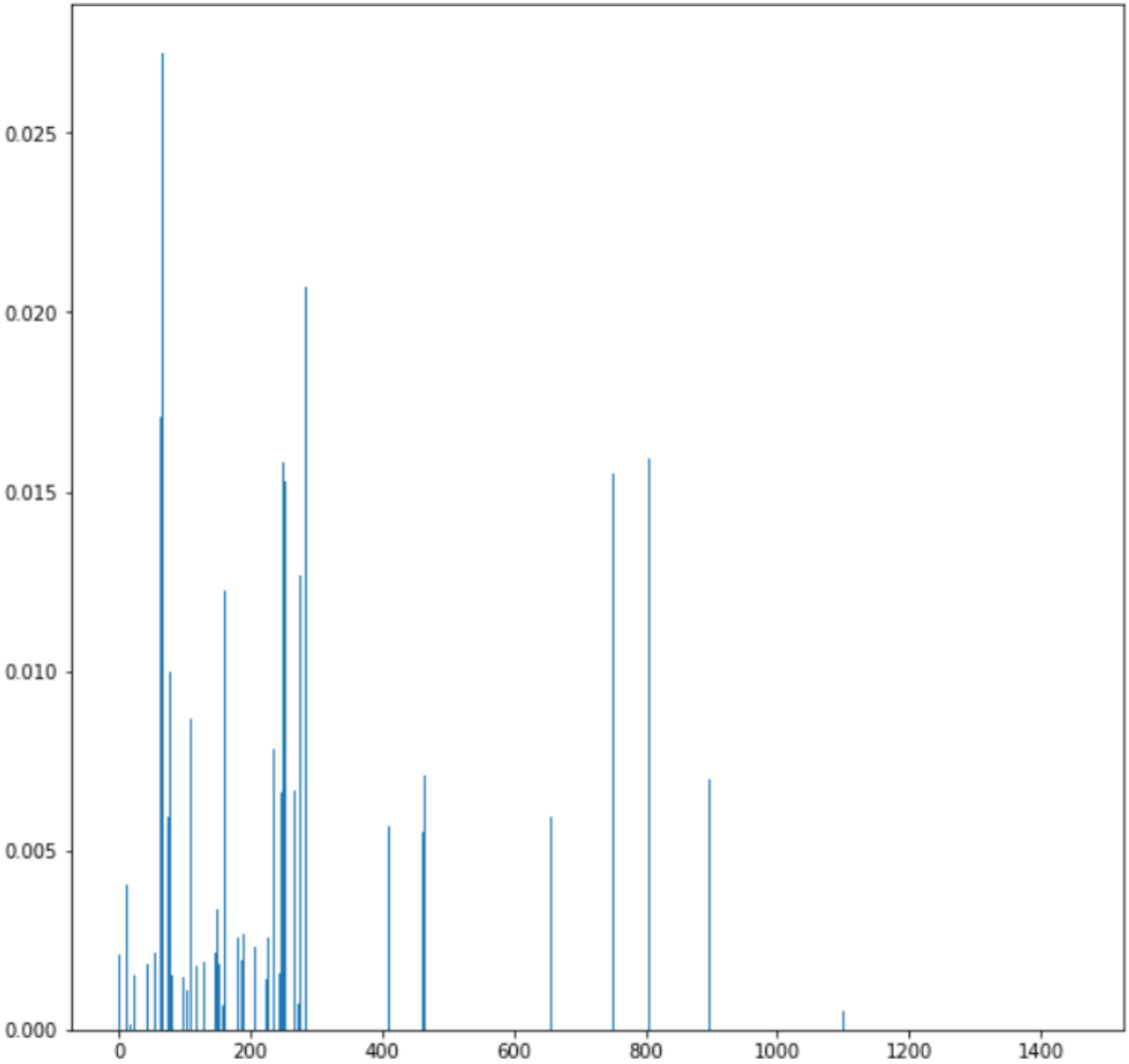
Feature importance - 1 year XGBoost



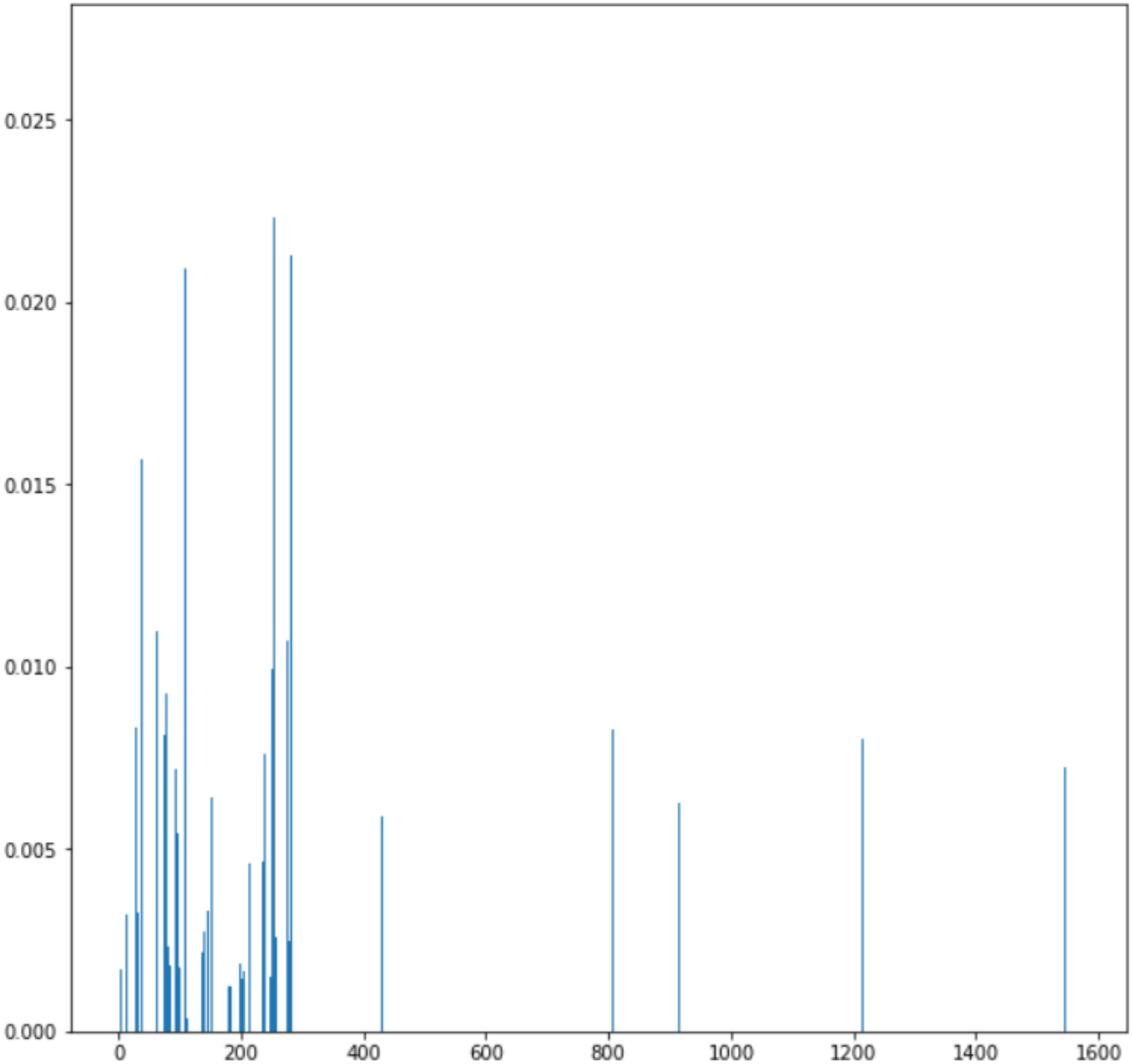
Feature importance - 3 year XGBoost



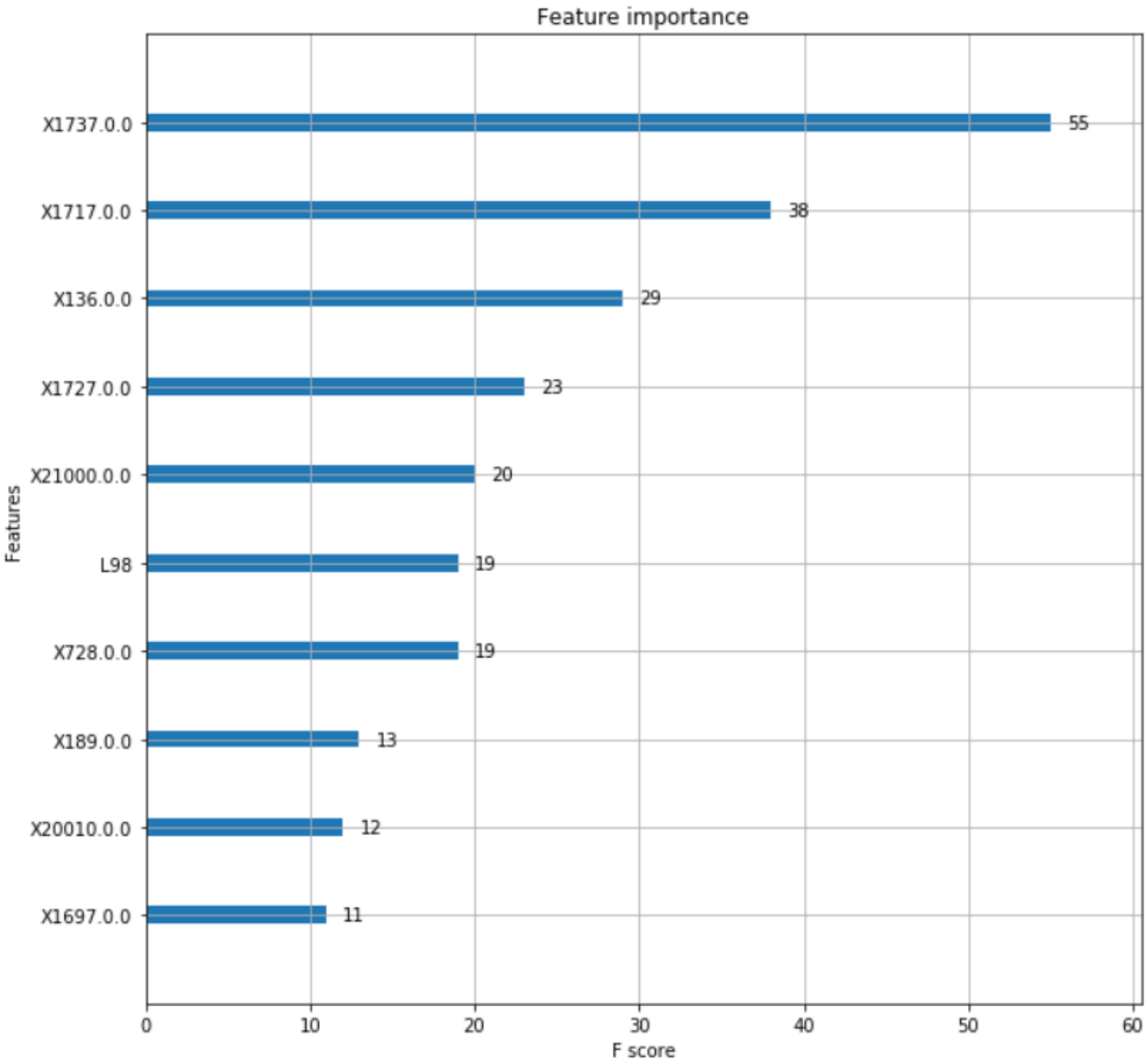
Feature importance - 5 year XGBoost



Feature importance - 10 year XGBoost



Feature importance - Top 10 features 1 year

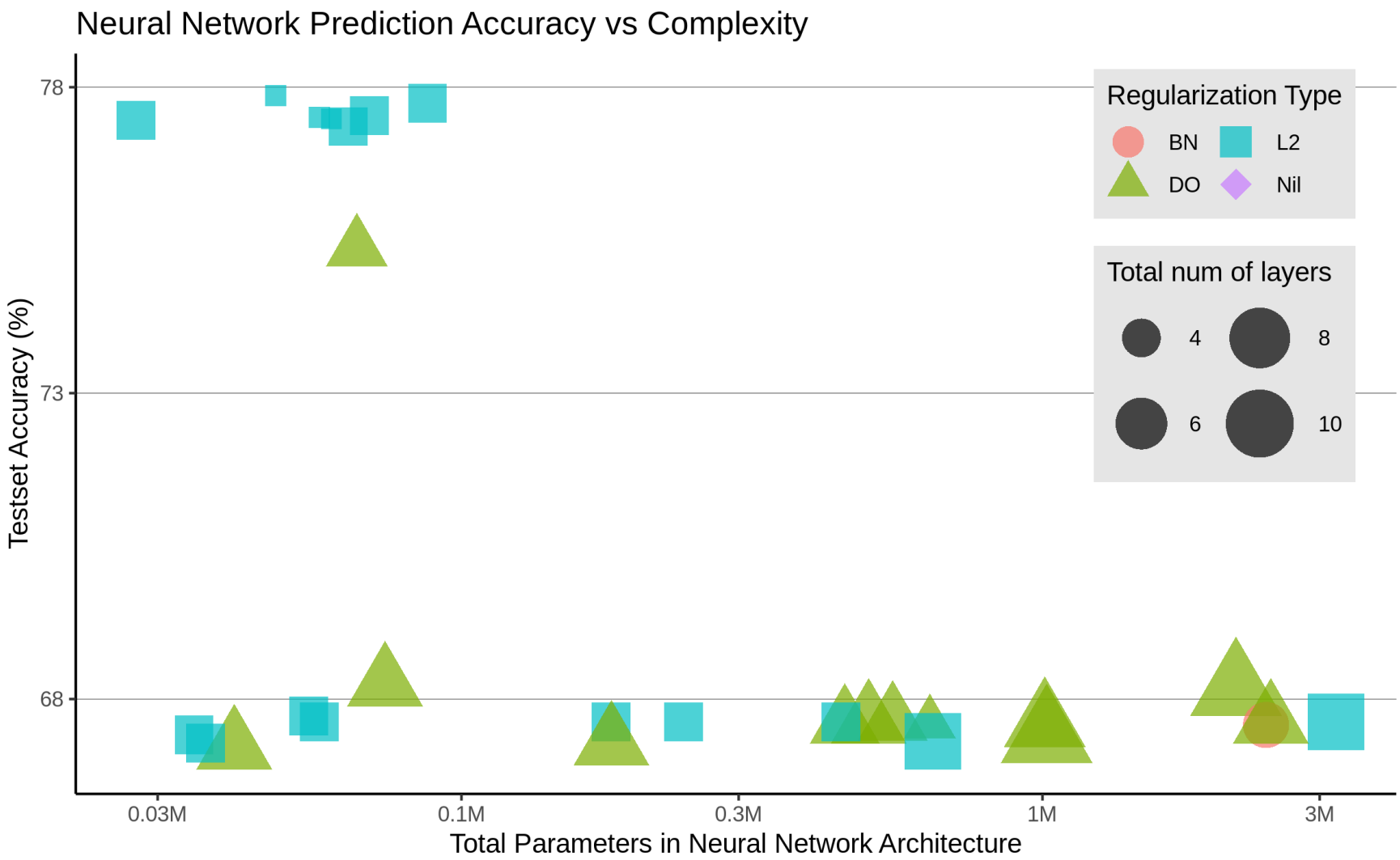


Feature importance

Top predictors for 1 year XGBoost

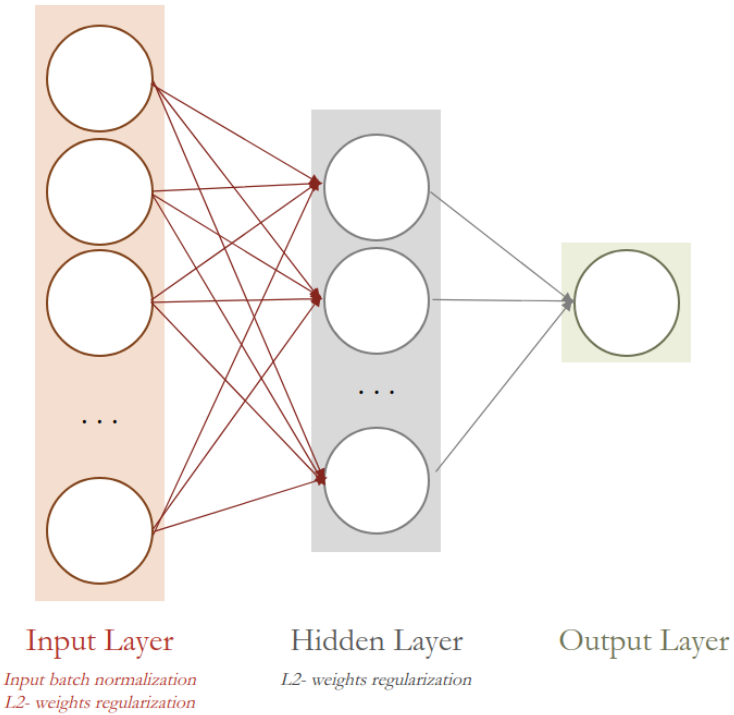
- Childhood sunburn occasions
- Skin colour
- Ease of skin tanning
- ICD-10-CM Code L98 - Other disorders of skin and subcutaneous tissue

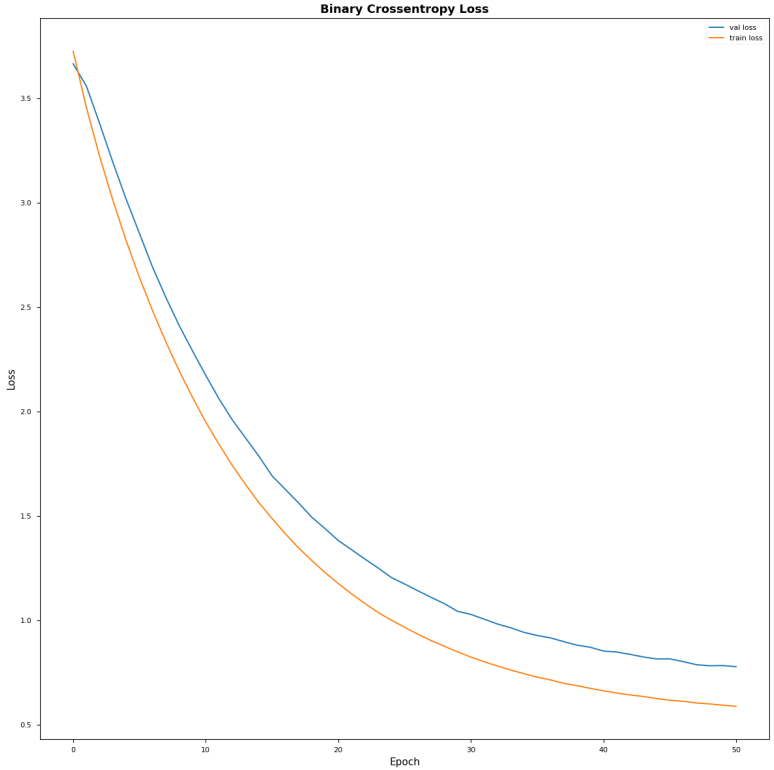
Neural Network



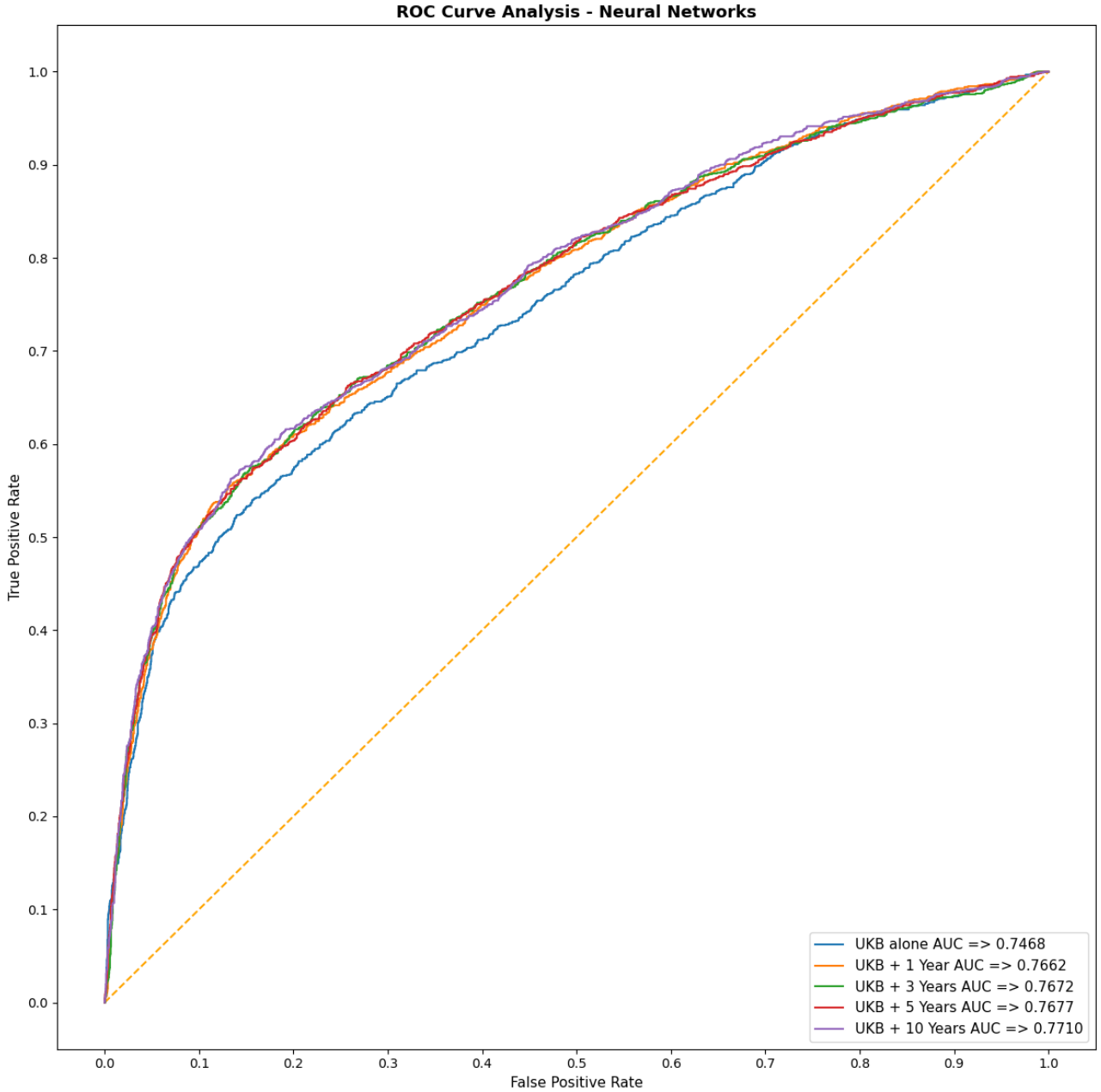
BN = Batch Normalization, DO = Dropout, L2 = L2-weights penalization

Neural Network





Neural Network



Discussion

- Building predicted models with a high degree of accuracy and low false negative rate is possible
- Tree-based models and neural networks outperformed logistic regression
  - Suggesting non-linear relationships between the predictors and outcome
- Time windows did not improve predictive performance after 1 year
  - This is consistent across all models that handle non-linear data
  - Suggests that information relevant to prediction (within the hospital admissions data) occurs up to 1 year before diagnosis
- The above implies a complex short-term relationship between malignant melanoma and its precursors

Limitations of Study

- Although we achieved good predictive accuracy
  - Lack in clinical utility from the short-term prediction of our model
  - Data required for models is not readily available for patients outside of research
- Our findings may not generalise to other populations
  - UKB is homogeneous
    - ★ Ethnicity
    - ★ Affluence
    - ★ Age

Implications for Further Work

- Shifting windows to earlier before diagnosis to aid in early diagnosis
- Link to primary care database may be more informative for time windows
  - Use less serious events that don't require hospitalisation
- Incorporation of genetic data as predictors may improve accuracy