Supplementary Material to:
Application of machine-learning algorithms to identify the key determinants of risk for HIV, Hepatitis C and Hepatitis B in primary care settings

Harrison Manley[a,i], Werner Leber[b,i], Kelvin Smith[b], Hamza Farooq[b], Manish Pareek[c], Rebecca F Baggaley[d], Jane Anderson[b,e], Leo Loman[a], Chris Griffiths[f], John Robson[b], Jasmina Panovska-Griffiths[a,g,h,i]

[a] *UK Health Security Agency, London, United Kingdom*
[b] *Queen Mary University London, London, United Kingdom*
[c] *University of Leicester, Leicester, United Kingdom*
[d] *University College London, London, United Kingdom*
[e] *Homerton Healthcare NHS Foundation Trust, London, United Kingdom*
[f] *Nuffield Department of Primary Care, University of Oxford, Oxford, United Kingdom*
[g] *The Pandemic Sciences Institute, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom*
[h] *The Queen's College, University of Oxford, Oxford, United Kingdom*
[i] *these authors contributed equally*

Description of risk factors

From the GP EHR data we extracted a number of features across the three BBVs. In Table S1 we describe the risk factors that were used as features within the machine-learning models across different BBVs in this study. In Table S2 we show the number of patients with positive tests but no risk factors across the three BBVs.

Target BBV	Datasets	Risk Factors
HIV	BBV Risks, HIV Associated conditions, AIDs defining conditions	Liver disease, Other substance use, Migrant, Cocaine use, Blood or transplant, Imprisonment, MSM, Piercing, Tattoo, Bacterial pneumonia, skin eruption, sborrheic darmatitis, lymphadenopathy, chronic kidney disease, psoriasis, herpes zoster, oral candidiasis, peripheral nerve disease, anogenital warts, abnormal weight loss, molluscum contagiosum, infectious mononucleosis, neutropenia, community acquired pneumonia, thrombocytopenia, dysplasia cervix, pyrexia, mononeuritis, thrichomoniasis, gonorrhoea, primary lung neoplasm, malignant lymphoma, multiple sclerosis, chronic diarrhoea, viral hepatitis A, syphilis, leuokpenia, pancytopenia, anogenital herpes, guillaine barre, anal tumour, chlamydia trachomitis, oral leukoplakia, lymphogranuloma venereum, candidiasis disseminated, leishmaniasis, chancroid, Recurrent bacterial pneumonia, Tuberculosis, Candiasis oesophagus, Cervical Cancer, Anogenital Herpes, Cytomegalovirus, Karposi Sarcoma, Pnuemocystis, Cryptosporidium, Cryptococcosis, Salmonella septicaemia, Leukoencephalopathy, Toxoplasmosis, Histoplasmosis
HBV	BBV Risks, HBV risks	Liver disease, Other substance use, Migrant, Cocaine use, Blood or transplant, Imprisonment, MSM, Piercing, Tattoo, Anti-hepatitis B immunoglobulin given, At risk of hepatitis B infection, hepatitis B occupational risk, Hepatitis B contact, Hepatitis B immunisation recommended, Hepatitis B screening required, mother Hepatitis B positive, at risk of BBV infection, viral hepatitis contact, contact with and exposure to viral hepatitis.
HCV	BBV Risks	Liver disease, Other substance use, Migrant, Cocaine use, Blood or transplant, Imprisonment, MSM, Piercing, Tattoo

Table S1: All risk factors from risk datasets, by BBV target. The multiple BBV target model used all the risk factors listed above.

Data exploration

The age distribution of the data is shown in figure S1 as a histogram, with the age distribution from the census shown for the North-East London (NEL) boroughs. There is generally a good match between the age distribution in the data used and the NEL boroughs census data.

BBV	Risk factor group	Num positives with no risk factors from group
HIV	BBV risks data	7427 (79.54 %)
	HIV Associated conditions	6380 (68.33 %)
	AIDs defining condition	4918 (52.67%)
	All of the above	3198 (34.26 %)
HBV	BBV risks data	12,691 (69.49 %)
	HBV risks data	13,956 (76.42 %)
	All of the above	9834 (53.85 %)
HCV	BBV risks data	2973 (43.89 %)

Table S2: Table showing the number of patients with a positive test but no risk factors for each BBV, and the percentage of the positive cohort they make up in each case.

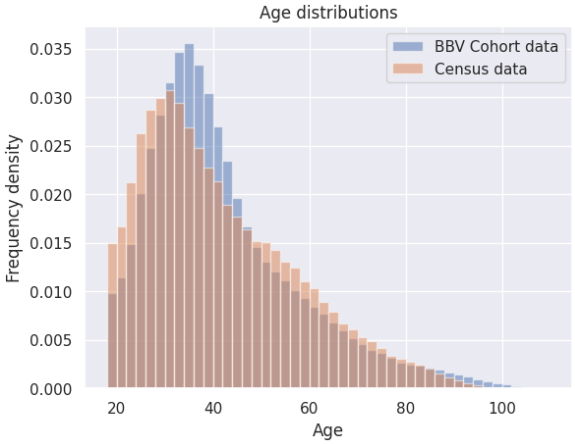

Figure S1: Comparison of the distribution of ages in the data and the 2021 census for the north east london boroughs in the study.

There is an approximately even split between the Male and Female in the full cohort (with unknown and other making up less than 0.01% of the data). Specifically, 49.807% are Female and 50.186% are Male.

Figure S2 compares the sex distribution between the general and BBV positive cohorts. Notably men comprise a higher proportion of the HIV, HBV and HCV cohorts, suggesting a higher likelihood of BBV positivity among males.

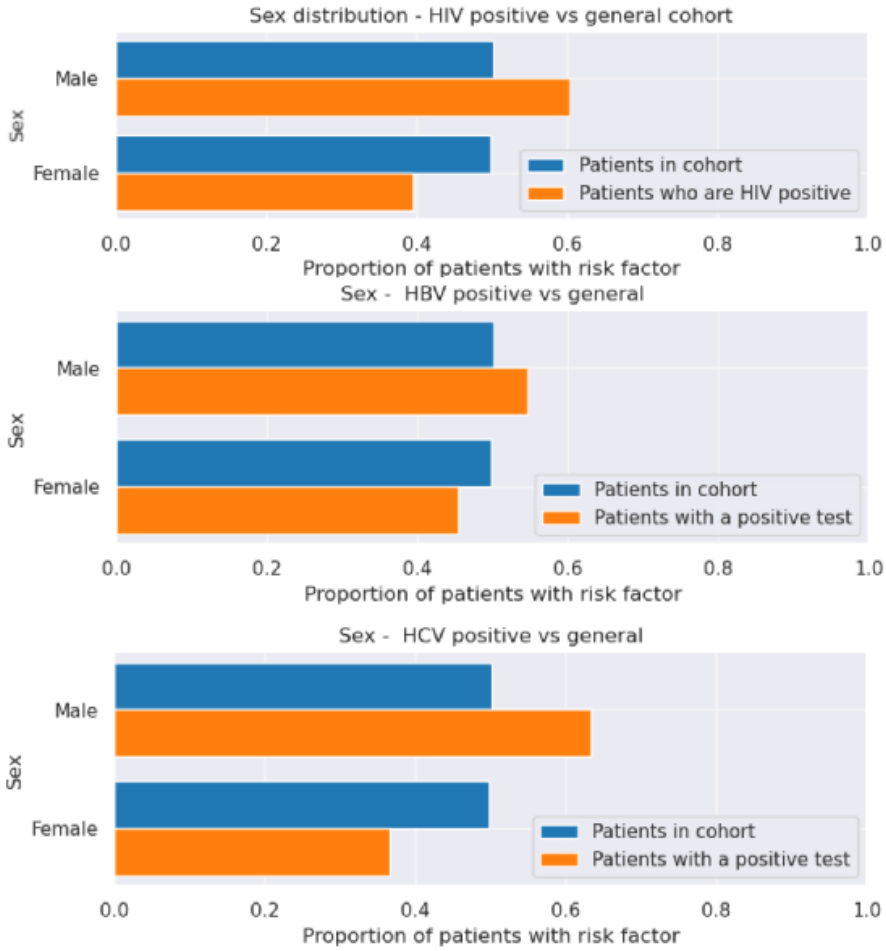

Figure S2: Distribution of sex for general cohort and BBV positive cohorts.

Figure S3 shows good agreement between our data and the NEL boroughs' census. The conversions from 18+1SDE to the ethnicity categories in the model are given in table S3.

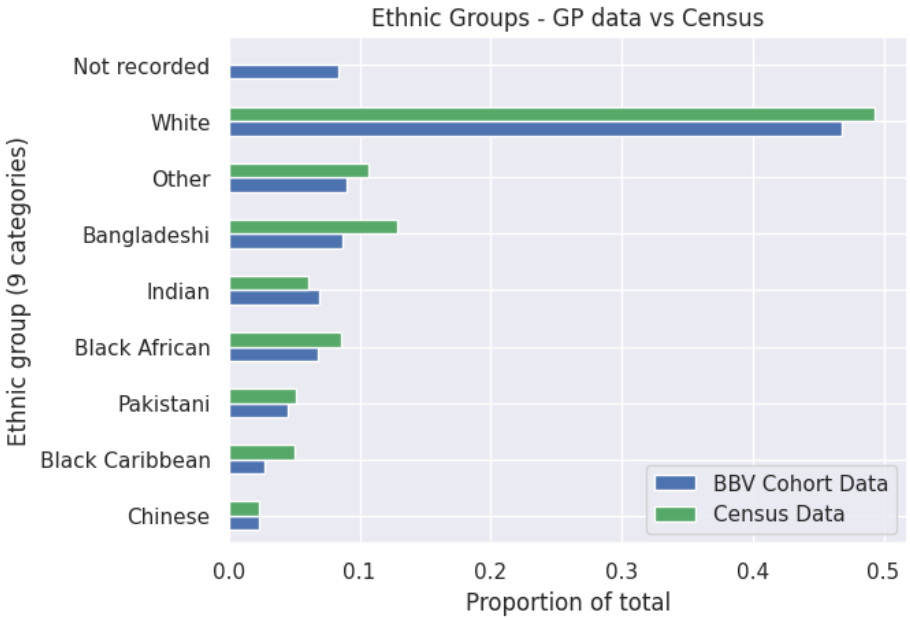

Figure S3: Ethnicities in our data, compared with the census data, shown as a horizontal bar plot.

SDE18+1 code	SDE18+1 name	8+1 Ethnic group name
W1	White British	White
W2	Irish	
W3	Gypsy and Irish Traveller	
W9	Any other White Background	
B1	Black Caribbean	Black Caribbean
B2	Black African	Black African
A1	Indian	Indian
A2	Pakistani	Pakistani
A3	Bangladeshi	Bangladeshi
A4	Chinese	Chinese
A9	Any other Asian background	Other Asian
M3	White and Asian	
M1	White and Black Caribbean	Other
M2	White and Black African	
O2	Arab	
O9	Any other ethnic group	
B9	Any other Black, Black British or Caribbean background	
UU	Unknown	Unknown
NS	Not Stated	

Table S3: Conversion table for 18+1 to 8+1 ethnicities.

Model performance: precision-recall curves

The precision recall (PR) curves for the individual BBV targets are shown in figure S4. We show that the BRFC and Logistic regression (LR) both comfortably outperform the chance level (shown by the dotted line on the plot). The PR Area-Under the Curve (PR AUC) is given in the plots. When predicting individual HIV positivity, LR was the best model using the PR AUC, followed by BRFC and then RUSBoost. When predicting individual HCV or HBV positivity, LR and BRFC were both equally good using the PR AUC, followed by RUSBoost. These results are very similar to those in the main text, where ROC AUC curves suggested LR and BRFC were notably better at predicting individual HIV, HCV or HBV positivity than RUSBoost.

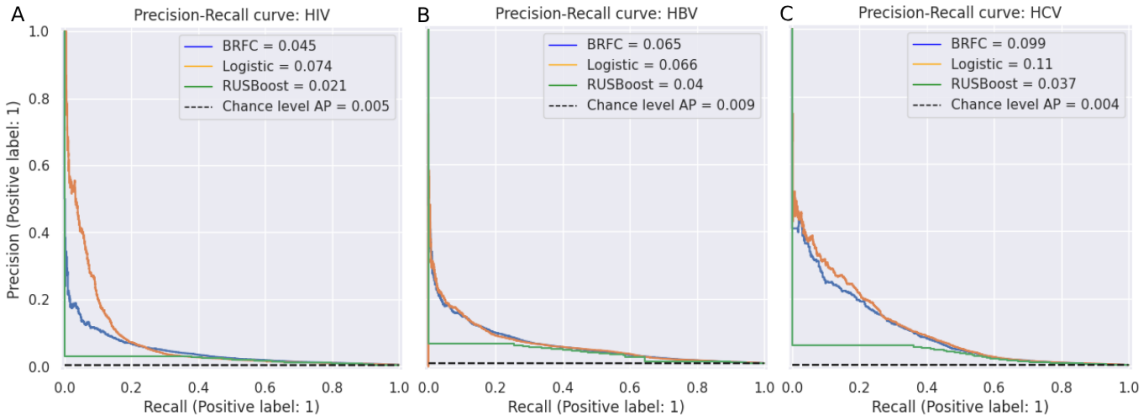

Figure S4: Precision-recall curves for HIV, HBC and HCV targets in panels A, B and C respectively.

The PR curves for the multiple BBV target are shown in figure S5. The LR is the best performing by this metric, and all three models comfortable outperform the chance level, shown by the dotted line on the plot. We also note that when predicting multiple positivity the PR AUC and the ROC AUC curves also show the same results: LR is the best predictive model for multiple HIV/HCV/HBV positivity using these AUC curves, followed by BRFC and then RUSBoost.

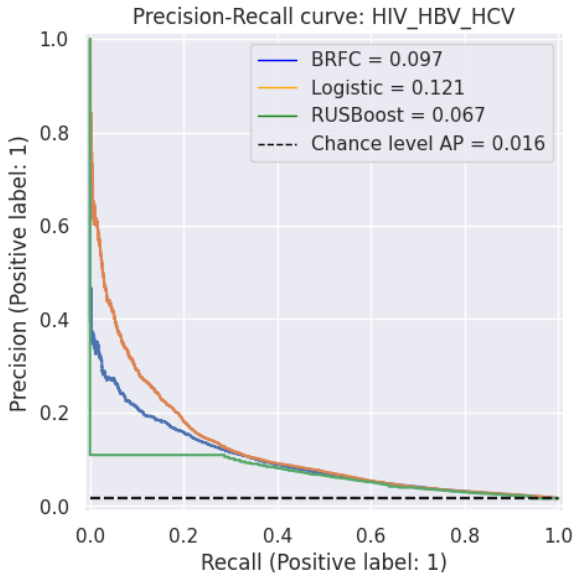

Figure S5: Precision-recall curves for the multiple BBV target models.

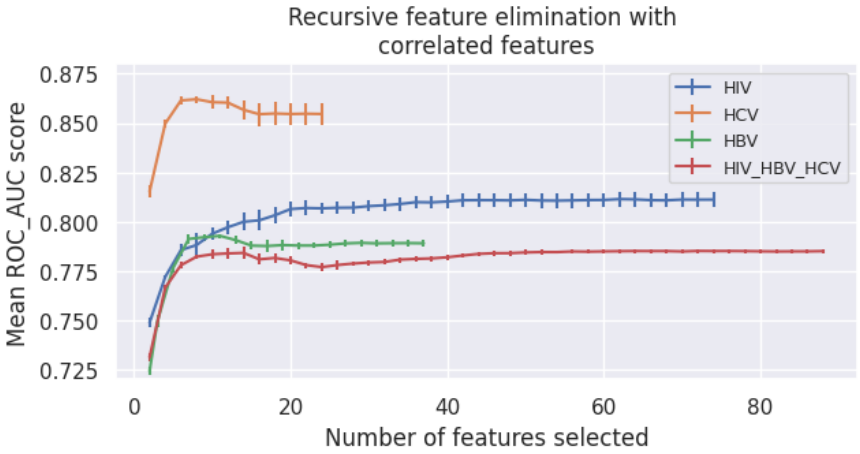

Figure S6: Plot showing the recursive feature elimination for the individual and multiple BBV target models, using the ROC AUC score, calculated as the mean and standard deviation of the ROC AUC value over a 4-fold cross validation.

Sensitivity Analysis

For the sensitivity analysis we looked at how the performance metrics changed when we removed patients with a negative test, patients recorded as deceased, and both negative testing and deceased patients. The results are very similar to the scores in the main text shown in Table 2. There were a slight difference in the optimal model when HBV positivity was predicted and specificity or sensitivity were used as accuracy metrics. These differences were, however, very small suggesting that inclusion of the negative testing and dead patients had minimal impact on the analysis.

In Tables S4, S5 and S6 we show the ROC AUC, sensitivity and specificity for each of the three models (LR, BRFC and RUSBoost) when predicting individual HIV, HCV and HBV positivity as well as when predicting multiple HIV/HVC/HBV positivity in three cases: when we exclude people with a negative test (Table S4), when we exclude patients recorded as deceased (Table S5) and when we exclude both people with a negative test and those patients recorded as deceased (Table S6). Across the three tables the results are very similar to the results in Table 2 of the main text. Notably, across all scenarios BRFC was the optimal model when predicting individual HIV positivity by ROC AUC, LR when using sensitivity and BRFC when using specificity as an accuracy metric. When predicting HCV positivity, LR or BRFC were similar and better models when using ROC AUC, LR when using sensitivity and BRFC when using specificity as an accuracy metric. When predicting HBV positivity, LR or BRFC were also similar and better models when using ROC AUC as an accuracy metric. When predicting HBV positivity and using sensitivity as an accuracy metric, LR was optimal when the data on patients that tested negative or deceased were included (in Table 2 of the main text) but BRFC was slightly better when these patients' data were excluded. Also, when predicting HBV positivity with specificity, RUSBoost was the best model when the data on patients that tested negative or deceased were included (in Table 2 of the main text), but LR was the best model when both those testing negative or deceased were excluded.

When determining the model that was best at predicting multiple HIV/HCV/HBV positivity the results in the main Figure 2 and Figures S4-S6 agree: LR was the best model when using ROC AUC, BRFC when using sensitivity and RUSBoost when using specificity as an accuracy metric.

BBV	Model	ROC AUC	Sensitivity	Specificity
HIV	BRFC	0.814 [0.809, 0.819]	0.715 [0.688, 0.743]	0.759 [0.732, 0.787]
	Logistic	0.799 [0.793, 0.805]	0.731 [0.715, 0.748]	0.718 [0.701, 0.734]
	RUSBoost	0.800 [0.797, 0.803]	0.727 [0.693, 0.762]	0.730 [0.695, 0.764]
HBV	BRFC	0.794 [0.792, 0.796]	0.681 [0.661, 0.702]	0.770 [0.750, 0.791]
	Logistic	0.794 [0.788, 0.800]	0.650 [0.617, 0.683]	0.793 [0.760, 0.826]
	RUSBoost	0.775 [0.769, 0.781]	0.639 [0.615, 0.663]	0.794 [0.770, 0.818]
HCV	BRFC	0.857 [0.854, 0.860]	0.748 [0.735, 0.760]	0.804 [0.791, 0.816]
	Logistic	0.862 [0.856, 0.868]	0.768 [0.734, 0.802]	0.781 [0.747, 0.815]
	RUSBoost	0.846 [0.841, 0.851]	0.725 [0.699, 0.750]	0.815 [0.789, 0.841]
Multiple BBV	BRFC	0.787 [0.781, 0.793]	0.695 [0.684, 0.707]	0.740 [0.728, 0.752]
	Logistic	0.789 [0.787, 0.791]	0.677 [0.669, 0.685]	0.747 [0.739, 0.755]
	RUSBoost	0.769 [0.763, 0.775]	0.658 [0.649, 0.667]	0.758 [0.749, 0.766]

Table S4: The ROC AUC, sensitivity and specificity scores for different models when predicting individual HIV, HCV or HBV or multiple HIV/HCV/HBV positivity, trained and tested on data without patients who had a negative test result. The values shown are the mean scores over 4 folds of the data using k-fold cross validation, with the 95% C.I. calculated from the 4 AUC values and the standard deviation, and given in the square brackets.

BBV	Model	ROC AUC	Sensitivity	Specificity
HIV	BRFC	0.811 [0.806, 0.816]	0.713 [0.674, 0.752]	0.757 [0.718, 0.796]
	Logistic	0.801 [0.795, 0.807]	0.725 [0.705, 0.745]	0.728 [0.708, 0.749]
	RUSBoost	0.795 [0.784, 0.806]	0.716 [0.642, 0.790]	0.734 [0.659, 0.808]
HBV	BRFC	0.790 [0.787, 0.793]	0.683 [0.662, 0.703]	0.764 [0.743, 0.785]
	Logistic	0.792 [0.787, 0.797]	0.673 [0.640, 0.706]	0.765 [0.732, 0.799]
	RUSBoost	0.768 [0.754, 0.782]	0.645 [0.627, 0.663]	0.779 [0.761, 0.797]
HCV	BRFC	0.854 [0.851, 0.857]	0.737 [0.727, 0.746]	0.811 [0.802, 0.821]
	Logistic	0.861 [0.851, 0.871]	0.762 [0.738, 0.785]	0.788 [0.764, 0.811]
	RUSBoost	0.841 [0.833, 0.849]	0.732 [0.703, 0.760]	0.806 [0.777, 0.834]
Multiple BBV	BRFC	0.785 [0.783, 0.787]	0.697 [0.685, 0.710]	0.734 [0.721, 0.746]
	Logistic	0.790 [0.784, 0.796]	0.698 [0.675, 0.720]	0.728 [0.706, 0.751]
	RUSBoost	0.769 [0.764, 0.774]	0.638 [0.612, 0.664]	0.774 [0.748, 0.800]

Table S5: The ROC AUC, sensitivity and specificity scores for different models when predicting individual HIV, HCV or HBV or multiple HIV/HCV/HBV positivity, trained and tested on data without patients recorded as deceased. The values shown are the mean scores over 4 folds of the data using k-fold cross validation, with the 95% C.I. calculated from the 4 AUC values and the standard deviation, and given in the square brackets.

BBV	Model	ROC AUC	Sensitivity	Specificity
HIV	BRFC	0.810 [0.800, 0.820]	0.723 [0.704, 0.743]	0.739 [0.720, 0.759]
	Logistic	0.802 [0.797, 0.807]	0.724 [0.705, 0.743]	0.729 [0.711, 0.748]
	RUSBoost	0.796 [0.788, 0.804]	0.733 [0.719, 0.747]	0.721 [0.706, 0.735]
HBV	BRFC	0.792 [0.787, 0.797]	0.671 [0.661, 0.682]	0.776 [0.766, 0.787]
	Logistic	0.796 [0.794, 0.798]	0.641 [0.624, 0.659]	0.808 [0.791, 0.826]
	RUSBoost	0.775 [0.773, 0.777]	0.650 [0.641, 0.658]	0.779 [0.770, 0.787]
HCV	BRFC	0.854 [0.852, 0.856]	0.731 [0.705, 0.756]	0.821 [0.795, 0.846]
	Logistic	0.861 [0.853, 0.869]	0.766 [0.757, 0.776]	0.783 [0.774, 0.792]
	RUSBoost	0.842 [0.829, 0.855]	0.730 [0.714, 0.747]	0.805 [0.789, 0.822]
Multiple BBV	BRFC	0.774 [0.772, 0.776]	0.675 [0.655, 0.694]	0.751 [0.732, 0.770]
	Logistic	0.779 [0.777, 0.781]	0.660 [0.649, 0.671]	0.767 [0.755, 0.778]
	RUSBoost	0.738 [0.735, 0.741]	0.644 [0.640, 0.649]	0.771 [0.767, 0.776]

Table S6: The ROC AUC, sensitivity and specificity scores for different models when predicting individual HIV, HCV or HBV or multiple HIV/HCV/HBV positivity, trained and tested on data without patients recorded as deceased and without patients who have a negative test result. The values shown are the mean scores over 4 folds of the data using k-fold cross validation, with the 95% C.I. calculated from the 4 AUC values and the standard deviation, and given in the square brackets.