Early prediction of prostate cancer risk in younger men using polygenic risk scores and electronic health records

Amita Varma
Dascena

Jenish Maharjan
Dascena

Anurag Garikipati
Dascena

Myrna Hurtado (MHurtado@dascena.com)
Dascena

Sepideh Shokouhi
Dascena

Qingqing Mao
Dascena

Research Article

Keywords: prostate cancer, polygenic risk score, machine learning, algorithms, decision support

Posted Date: June 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1769754/v1

Abstract

Background

Prostate cancer (PCa) screening is not routinely conducted in men 55 and younger, although this age group accounts for more than 10% of cases. Polygenic risk scores (PRSs) and patient data applied towards early prediction of PCa may lead to earlier interventions and increased survival. We have developed machine learning models to predict PCa risk in men 55 and under using PRSs combined with patient data.

Methods

We conducted a retrospective study on 91,106 male patients aged 35 to 55 using the UK Biobank database. Five gradient boosting models were developed and validated utilizing routine screening data, PRSs, additional clinical data, or combinations of the three.

Results

Combinations of PRSs and patient data outperformed models that utilized PRS or patient data only, and the highest performing models achieved an area under the receiver operating characteristic curve of 0.788. Our models demonstrated a substantially lower false positive rate (35.4%) in comparison to standard screening using prostate specific antigen (60–67%).

Conclusion

This study provides the first preliminary evidence for the use of PRSs with patient data in a machine learning algorithm for PCa risk prediction in men 55 and under for whom screening is not standard practice.

Introduction

Prostate cancer (PCa) is the second most common cancer in men and responsible for 375,000 deaths worldwide.(1) Although it presents an indolent clinical course, PCa still remains a major health burden with mortality rates expected to rise 1.05% by 2040.(2) PCa is generally asymptomatic in the early and later stages.(3, 4) Routine cancer screening can prevent future health complications by facilitating early detection and allowing for timely intervention. The most common screening methods for PCa are the digital rectal examination (DRE) and prostate-specific antigen (PSA) test. The largest conducted trial of DRE and PSA screening demonstrated the usefulness of screening with a subsequent risk reduction in PCa-related deaths of up to 49%.(5) However, there is controversy surrounding the effectiveness of PSA screening as false positive results, overdiagnosis, and overtreatment are associated with use of this screening tool.(6) In 2012, the United States Preventive Services Task Force issued a recommendation discouraging routine PCa screening in men regardless of risk factors, causing high-grade cases to increase by 11.3%.(4) Further efforts are warranted to improve current PCa initial screening approaches and methods.

Screening is generally recommended for men aged 55 and older, as the majority of PCa cases are diagnosed in older men. Although the average age of PCa diagnosis is 66, with the highest incidence seen in those older than 65,(7) more than 10% of cases occur in men 55 and younger(8) and current research indicates that younger men diagnosed with high-grade PCa have an overall poorer prognosis.(9) Developing an accurate screening tool to predict the risk of PCa for patients younger than the standard screening age would therefore allow for earlier identification of those younger patients at risk and potentially reduce the public health burden.

The high heritability of PCa(10) demonstrates that genetic factors play a considerable role in its development. Several genome-wide association studies have identified over 170 single nucleotide polymorphisms (SNPs) that are associated with an increased risk of PCa.(11) These genetic variants can be combined to determine an individual's polygenic risk score (PRS), and PRSs have been demonstrated to have a large clinical utility potential for numerous diseases, including PCa.(12)

PCa is also associated with additional known risk factors, such as age and ethnicity,(13) that can be routinely entered into electronic health records. PRSs along with patient data may be used for earlier and more accurate predictions of PCa, leading to earlier interventions, increased survival, and reduced healthcare costs. We have developed and validated machine learning (ML) models to predict PCa diagnosis specifically in younger men (age ≤ 55) based on PRS and relevant patient data. This risk assessment screening method is not contingent on the use of PSA or DRE results.

Methods

Data source

Data from 502,460 participants in the UK Biobank (UKBB) were analyzed in this retrospective study. UKBB is a longitudinal electronic health record repository that incorporates clinical and genetic data. Patient data from hospitals and UKBB assessment centres between January 2007 to June 2020 were used in this study. Prior to use, passive patient data gathering and de-identification was conducted in compliance with the Health Insurance Portability and Accountability Act. The use of de-identified retrospective data is classified as a non-human subject study and exempt from Institutional Review Board approval.

Cohort definition and Gold Standard

We included all male UKBB participants aged 35 to 55 who had genotypic data. The gold standard labels for a positive PCa diagnosis were defined using data from two fields: International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10) diagnoses, and self-reported cancer code. The

ICD-10 code used to define PCa was C61 - Malignant neoplasm of prostate, and the self-reported cancer code was 1044 - Prostate Cancer. Any patient fitting either of these two criteria was labeled as positive. Those that had a PCa diagnosis prior to this visit were excluded. All other patients were considered negative cases.

Genetic Data and PRS

The PRSs were created using the PRSice tool (https://www.prsice.info/quick_start/).(14) The polygenic score (PGS) weights found on the PGS Catalog website (https://www.pgscatalog.org/score/PGS000333/)(15) were used to generate the PRSs for every participant. The genome wide association study for this set of weights was performed on a cohort of European ancestry(11) for PCa among other traits. This set of weights were then trained and validated on the FINRISK biobank cohort.(15) The study reported weights and summary statistics of 6,606,785 SNPs. The UKBB cohort had a total of 784,256 variants out of which 541,268 variants overlapped with the variants reported by the study.

Machine Learning Algorithm, Input Features, and Prediction Models

We used XGBoost, a gradient boosting algorithm(16), implemented in Python. This algorithm was chosen because it allows the analysis of contributions of individual features to the algorithm results. Five models were developed: 1) PRSs only, 2) Features I: utilizing only age, father's history, sibling history and ethnicity, 3) PRSs + Features I, 4) Minimal Features: age, father's history and body mass index (BMI), and 5) PRSs + Minimal Features. Four additional models were investigated: 1) Features II: Features I + BMI, smoking status, glycated hemoglobin (HbA1c), C-reactive protein, and insulin-like growth factor 1, 2) PRS + Features II, 3) Features III: Features II + number of same sex partners, diabetes diagnosis, and diabetes medication and 4) PRS + Features III. Apart from HbA1c, these supplementary features did not improve model performance and are not the main models of focus in our study. We partitioned the dataset into training (60%), validation (20%), and hold-out testing (20%) splits prior to training of the model. The validation set was used during training to validate the model performance. The hold out test set is not seen by the model during the training or validation phase. Results are reported for the hold out test set. The model was trained to predict PCa up to 11 years, the maximum time between the patient's visit to the health facility used for training and the first diagnosis of PCa in the data set. Missing values in the continuous features were filled as null, and missing values in categorical features were filled with the appropriate data code for "Unknown" or "No response". Family history features are binary features, created by checking the presence of the code for PCa in the appropriate columns for father and sibling history. The hyperparameters of the model were tuned on the validation set using a 3-fold grid search cross-validation approach. The hyperparameters that were tuned were *eta* (learning rate), *gamma* (minimum loss reduction to split), and *lambda* (L2 regularization term). The number of estimators was fixed to be 100, and the maximum depth to be 6.

Statistical analysis

The performance of each model was evaluated on the 20% hold-out test set with respect to the area under the receiver operating characteristic (AUROC), sensitivity, specificity, diagnostic odds ratio, and positive and negative likelihood ratios. The threshold for predicting labels was calculated by setting the minimum sensitivity value to 0.800. 95% confidence intervals for these metrics were constructed using 1000 bootstrapped samples. We conducted a SHapely Additive ExPlanations (SHAP) plot(17) to evaluate feature importance.

Results

Subject characteristics

There were 502,460 UKBB participants before exclusion of any patients. The number of patients was 229,106 after excluding female patients. After exclusion of male patients over the age of 55 and those with a prior PCa diagnosis, a total of 91,106 men were included in the study: 90,419 control participants and 687 participants with a PCa diagnosis. Figure 1 represents the attrition chart. Table 1 summarizes the patient characteristics for positive cases (participants with PCa diagnosis) and controls (participants with no PCa diagnosis). Group differences were calculated with Fisher's exact test. Age ($p < 0.0001$) and black ethnicity ($p < 0.0001$), showed significant group differences. Table 2 provides the input features for the five ML models: PRS only, Features I (age, father's history, sibling history and ethnicity), PRS + Features I, Minimal Features (age, father's history and BMI), PRS + Minimal Features. Supplementary Table 1 lists the input variables for the four additional ML models: Features II, PRS + Features II, Features III, and PRS + Features III.

Table 1
Demographic data and other patient characteristics for individuals with and without prostate cancer (PCa) included in the hold out test set.

Demographics		*With PCa (n = 687)*	*Without PCa (n = 90,419)*	*P-value*
Age	35–45	29 (4.2%)	23334 (25.8%)	< 0.0001
	45–55	658 (95.8%)	67085 (74.2%)	< 0.0001
Ethnicity	White	624 (90.8%)	82384 (91.1%)	0.978
	Black	37 (5.4%)	2239 (2.5%)	< 0.0001
	Asian	13 (1.9%)	3377 (3.7%)	0.010
	Mixed	7 (1.0%)	683 (0.8%)	0.375
	Other	6 (0.9%)	1736 (1.9%)	0.048
BMI	Mean	27.53	27.76	0.186
	Range (Min-Max)	16.75–43.16	14.87–63.44	-
Smoking Status	Current smoker	95 (13.8%)	13886 (15.4%)	0.369
	Previous smoker	185 (26.9%)	24721 (27.3%)	0.901
	Never smoker	406 (59.1%)	51413 (56.9%)	0.549
	Did not answer	1 (0.1%)	399 (0.4%)	0.381

Table 2
Input variables for the five machine learning (ML) models: PRS only, Features I, PRS + Features I, Minimal Features and PRS + Minimal Features.

	Machine Learning Model				
Input Variables	*1) PRS only*	*2) Features I*	*3) PRS + Features I*	*4) Minimal Features*	*5) PRS + Minimal Features*
Genetic	PRS		PRS		PRS
Demographics		Age	Age	Age	Age
		Father's history	Father's history	Father's history	Father's history
		Sibling history	Sibling history		
		Ethnicity	Ethnicity		
Clinical Measurements				BMI	BMI

ML Algorithm Performance

Figure 2 shows the ROC curves for the five models. Table 3 summarizes the performance metrics of all five models and includes the AUROC, sensitivity, specificity, diagnostic odds ratio, false positive rates, and positive and negative likelihood ratios. The PRS + Features I and PRS + Minimal Features models' performance were comparable and demonstrated the highest AUROCs, 0.788 (95% CI = 0.758–0.819) and 0.788 (95% CI = 0.757–0.820), respectively. At a sensitivity of 0.800, the PRS + Features I model demonstrated a specificity of 0.629 and the PRS + Minimal Features model a specificity of 0.646. The PRS + Minimal Features model had a false positive rate of 35.4%. The performance metrics of the four additional models (Features II, PRS + Features II, Features III, and PRS + Features III) are presented in Supplementary Table 2.

Table 3
Performance metrics of all five machine learning algorithm models. Abbreviations: Area under the receiver operating characteristic (AUROC); diagnostic odds ratio (DOR); false positive rate (FPR), likelihood ratio positive (LR+), likelihood ratio negative (LR-), polygenic risk score (PRS).

	Only PRS scores	*Features I only*	*PRS + Features I*	*Minimal features only*	*PRS + Minimal features*
AUROC	0.669 (0.634, 0.708)	0.750 (0.714, 0.781)	0.788 (0.758, 0.819)	0.751 (0.714, 0.784)	0.788 (0.757, 0.820)
Sensitivity	0.920 (0.877, 0.963)	0.807 (0.743, 0.870)	0.800 (0.736, 0.864)	0.807 (0.743, 0.870)	0.800 (0.736, 0.864)
Specificity	0.295 (0.289, 0.302)	0.552 (0.545, 0.560)	0.629 (0.622, 0.636)	0.563 (0.556, 0.570)	0.646 (0.639, 0.653)
DOR	4.819 (4.229, 5.410)	5.151 (4.744, 5.557)	6.783 (6.382, 7.184)	5.377 (4.971, 5.784)	7.299 (6.897, 7.700)
LR+	1.306 (1.244, 1.370)	1.802 (1.664, 1.953)	2.157 (1.986, 2.341)	1.846 (1.704, 2.000)	2.260 (2.081, 2.454)
LR-	0.271 (0.157, 0.466)	0.350 (0.252, 0.485)	0.318 (0.231, 0.438)	0.343 (0.248, 0.476)	0.310 (0.225, 0.426)
FPR	0.705 (0.698, 0.711)	0.448 (0.440, 0.455)	0.371 (0.364, 0.378)	0.437 (0.430, 0.444)	0.354 (0.347, 0.361)

Feature Importance

The SHAP plot (Fig. 3) shows the features with the highest contribution to the XGB results for the PRS + Minimal Features model. Age, PRS, and father's PCa history were identified as the top features having a positive association with PCa risk, whereas higher BMI was associated with lower risk. Sibling history and ethnicity were also identified as high-importance predictors in the Features I model and HbA1c in the Features II model; however, inclusion of PRS resulted in a sharp decrease in their feature importance. Supplementary Fig. 1 presents the SHAP plots of the Features I, PRSs + Features I, Features II, PRS + Features II model, and Minimal Features ML models.

Discussion

Summary of the Study

This is the first study demonstrating the utility of ML algorithms for PCa risk assessment in younger men who have not reached the recommended age for routine PCa screening. We achieved the same accuracy as the PRS + Features I model with fewer inputs (PRS + Minimal Features) and successfully created a risk assessment tool for identifying high-risk individuals among men aged 55 and younger. PCa incidence in men of this age group has been steadily increasing over the last few decades and is expected to continue rising. Younger men with high grade PCa have a significantly diminished overall survival and disease-specific survival compared to older men.(18, 19) Our ML-based prediction model may aid in the early detection of PCa in young at-risk individuals to prompt further examination and provide an opportunity for early treatment and prevention options.

Significance and impact of PRS in PCa prediction models

There are notable biological differences between early and late onset PCa, which can have significant clinical implications.(20) The early onset of PCa in younger men is thought to be largely attributed to genetic factors.(21) This is consistent with the presence of PRS and paternal PCa history among the top features of the ML prediction models. Although familial history is a known PCa risk factor, using genetic data in the form of a PRS can provide a more objective risk profile that is not contingent upon accurate information from an individual's family members.(22) As research continues to identify new PCa susceptibility loci,(11) we expect that future ML models will incorporate improved PRSs in parallel with new genetic discoveries.

Importance of other features

When PRS was not included in our model, ethnicity was among the significant features. PCa is known to be disproportionately higher in black men,(23) which is consistent with the significant overrepresentation of black men with PCa cases in comparison to other racial groups in our study. Sibling history and HbA1c were additional model features determined to be of high importance, although the addition of PRS to the ML model substantially reduced their importance as predictors. Incorporation of PRS diminished the need for these additional features and maintained high accuracy in our Minimal Features model. Interaction and overlapping effects between PRS and race/ethnicity, sibling history and HbA1c were not explored in this study and warrant further investigation.

Age was identified as the most important feature in predicting PCa, in agreement with established literature.(7) Addition of other clinical data (BMI) also slightly improved performance metrics. BMI has been reported to influence PCa aggressiveness, however the mechanisms are not yet known.(24–26) We observed a negative association between BMI and PCa risk. Similar findings were reported by Giovannuci et al. who found a lower risk of PCa in men with higher BMI only if they were younger (< 60 years old) or had a family history of prostate cancer, and attributed their findings to the complex relationship between obesity and various hormones.(27)

Comparison with other models

Previous studies in older men (ranging from 55–80 years) have reported that the inclusion of PRS data improves the performance of different PCa prediction models.(28–31) Oh et al. identified several ethnicity-specific SNPs with moderate predictive performance (an AUROC of 0.637) in men 60 years and older.(30) Aly et al. developed a baseline model in men 80 years and under based on age, PSA levels and familial history and determined that the inclusion of PRS improved performance metrics (AUROC of 0.64 to 0.67) as well as reduced the number of required diagnostic biopsies.(31) Our models did not rely on the use of PSA tests, which can be associated with false positive results leading to biopsy complications,(32) particularly in younger men where poorly differentiated adenocarcinoma may impact the accuracy of PSA as PCa risk predictor.(9) Our model demonstrated a substantial reduction in false positive rates for ML PCa screening compared to using PSA serum tests: 35.4% for our PRS + Minimal Features model in men aged 35–55, versus 60–67% for PSA screening in men aged 55–71.(33) This was also a considerable improvement from the 70.5% false positive rate of our PRS only model, which is comparable to current false positive rates of PSA screening in older men. Our ML models' combination use of genetic and patient data demonstrated increased accuracy in the identification of PCa risk in a younger cohort of men.

Study limitations

There are several limitations to this study. First, this work was conducted retrospectively, therefore we cannot determine how this model would perform in prospective clinical practice. Additionally, PRS studies that use cohorts of mainly European descent, as is the case with our dataset, may not be generalizable to other populations and may affect risk prediction accuracy for individuals of non-European ancestry.(34) Information on cancer aggressiveness (Gleason grades) was not available in our database and is an aspect to investigate in future work. Other PCa risk assessment tools(35, 36) require PSA or DRE measures, which were not available for direct comparison of risk assessment in this population. This work provides promising preliminary evidence for PCa risk evaluation in younger men and warrants future studies that should include validation of our ML algorithm in a prospective clinical setting and assessing how patient care and outcomes are affected.

Conclusion

ML algorithms which include PRS information and basic patient data can provide risk assessment for PCa in a young population not routinely screened. Efforts to identify men at risk in earlier age groups can help decrease the burden of PCa. Future work to support implementation of ML algorithms for PCa risk assessment of younger men in clinical practice is needed.

Declarations

Additional Information

Acknowledgments

We thank Misty M Attwood, PhD for her contribution to the research and development of this project.

Authors' contributions

AV designed the models and the computational framework, analyzed the data and contributed to the writing of the manuscript. JM designed the models and the computational framework and analyzed the data. MH and SS contributed to the writing and editing of the manuscript. AG and QM conceived of the project idea.

Ethics approval and consent to participate

Data were collected passively and de-identified in compliance with the Health Insurance Portability and Accountability Act, thus, this study was considered non-human subjects research and did not require Institutional Review Board approval.

Data availability

The data used in this study are available from UKBB and may be accessed by completing an application via https://www.ukbiobank.ac.uk/register-apply/. The ML algorithm code developed in this study is proprietary and not publicly available.

Competing interests.

The authors declare no competing interests.

Funding information

The authors received no specific funding for this work.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209–49.
2. Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin DM, Piñeros M, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer. 2019 Apr 15;144(8):1941–53.
3. Merriel SWD, Funston G, Hamilton W. Prostate Cancer in Primary Care. Adv Ther. 2018;35(9):1285–94.
4. Leslie SW, Soon-Sutton TL, Sajjad H, Siref LE. Prostate Cancer. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2021 [cited 2021 Sep 9]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK470550/
5. van den Bergh RCN, Loeb S, Roobol MJ. Impact of Early Diagnosis of Prostate Cancer on Survival Outcomes. Eur Urol Focus. 2015 Sep;1(2):137–46.
6. Palsdottir T, Nordstrom T, Karlsson A, Grönberg H, Clements M, Eklund M. The impact of different prostate-specific antigen (PSA) testing intervals on Gleason score at diagnosis and the risk of experiencing false-positive biopsy recommendations: a population-based cohort study. BMJ Open. 2019 Mar 30;9(3):e027958.
7. Rawla P. Epidemiology of Prostate Cancer. World J Oncol. 2019 Apr;10(2):63–89.
8. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. CA Cancer J Clin. 2012;62(1):10–29.
9. Gupta S, Gupta A, Saini AK, Majumder K, Sinha K, Chahal A. Prostate Cancer: How Young is too Young? Curr Urol. 2017 Jan;9(4):212–5.
10. Mucci LA, Hjelmborg JB, Harris JR, Czene K, Havelick DJ, Scheike T, et al. Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. JAMA. 2016 Jan 5;315(1):68–76.
11. Schumacher FR, Al Olama AA, Berndt SI, Benlloch S, Ahmed M, Saunders EJ, et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nat Genet. 2018 Jul;50(7):928–36.
12. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. Hum Mol Genet. 2019 Nov 21;28(R2):R133–42.
13. Leitzmann MF, Rohrmann S. Risk factors for the onset of prostatic cancer: age, location, and behavioral correlates. Clin Epidemiol. 2012 Jan 5;4:1–11.
14. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. GigaScience [Internet]. 2019 Jul 1 [cited 2021 Jul 29];8(7). Available from: https://doi.org/10.1093/gigascience/giz082
15. Mars N, Koskela JT, Ripatti P, Kiiskinen TTJ, Havulinna AS, Lindbohm JV, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nat Med. 2020 Apr;26(4):549–57.

16. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. San Francisco California USA: ACM; 2016 [cited 2021 May 24]. p. 785–94. Available from: https://dl.acm.org/doi/10.1145/2939672.2939785
17. Rodríguez-Pérez R, Bajorath J. Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. J Med Chem. 2020 Aug 27;63(16):8761–77.
18. Lin DW, Porter M, Montgomery B. Treatment and survival outcomes in young men diagnosed with prostate cancer: a population based cohort study. Cancer. 2009 Jul 1;115(13):2863–71.
19. Merrill RM, Bird JS. Effect of young age on prostate cancer survival: a population-based assessment (United States). Cancer Causes Control CCC. 2002 Jun;13(5):435–43.
20. Salinas CA, Tsodikov A, Ishak-Howard M, Cooney KA. Prostate Cancer in Young Men: An Important Clinical Entity. Nat Rev Urol. 2014 Jun;11(6):317–23.
21. Lange EM, Salinas CA, Zuhlke KA, Ray AM, Wang Y, Lu Y, et al. Early Onset Prostate Cancer Has A Significant Genetic Component. The Prostate. 2012 Feb 1;72(2):147–56.
22. Sun J, Na R, Hsu F-C, Zheng SL, Wiklund F, Condreay LD, et al. Genetic Score Is an Objective and Better Measurement of Inherited Risk of Prostate Cancer than Family History. Eur Urol. 2013 Mar;63(3):585–7.
23. Edwards BK, Howe HL, Ries LAG, Thun MJ, Rosenberg HM, Yancik R, et al. Annual report to the nation on the status of cancer, 1973–1999, featuring implications of age and aging on U.S. cancer burden. Cancer. 2002 May 15;94(10):2766–92.
24. Haque R, Van Den Eeden SK, Wallner L, Richert-Boe K, Kallakury B, Wang R, et al. Association of Body Mass Index and Prostate Cancer Mortality. Obes Res Clin Pract. 2014;8(4):e374–81.
25. Rodriguez C, Freedland SJ, Deka A, Jacobs EJ, McCullough ML, Patel AV, et al. Body Mass Index, Weight Change, and Risk of Prostate Cancer in the Cancer Prevention Study II Nutrition Cohort. Cancer Epidemiol Prev Biomark. 2007 Jan 1;16(1):63–9.
26. Lavalette C, Cordina Duverger E, Artaud F, Rébillard X, Lamy P, Trétarre B, et al. Body mass index trajectories and prostate cancer risk: Results from the EPICAP study. Cancer Med. 2020 Jul 8;9(17):6421–9.
27. Giovannucci E, Rimm EB, Liu Y, Leitzmann M, Wu K, Stampfer MJ, et al. Body Mass Index and Risk of Prostate Cancer in U.S. Health Professionals. JNCI J Natl Cancer Inst. 2003 Aug 20;95(16):1240–4.
28. Sipeky C, Talala KM, Tammela TLJ, Taari K, Auvinen A, Schleutker J. Prostate cancer risk prediction using a polygenic risk score. Sci Rep. 2020 Oct 13;10:17075.
29. Black MH, Li S, LaDuca H, Lo M, Chen J, Hoiness R, et al. Validation of a prostate cancer polygenic risk score. The Prostate. 2020 Nov 1;80(15):1314–21.
30. Oh JJ, Kim E, Woo E, Song SH, Kim JK, Lee H, et al. Evaluation of Polygenic Risk Scores for Prediction of Prostate Cancer in Korean Men. Front Oncol. 2020 Oct 22;10:583625.

31. Aly M, Wiklund F, Xu J, Isaacs WB, Eklund M, D'Amato M, et al. Polygenic Risk Score Improves Prostate Cancer Risk Prediction: Results from the Stockholm-1 Cohort Study. Eur Urol. 2011 Jul;60(1):21–8.
32. Fenton JJ, Weyrich MS, Durbin S, Liu Y, Bang H, Melnikow J. Prostate-Specific Antigen-Based Screening for Prostate Cancer: Evidence Report and Systematic Review for the US Preventive Services Task Force. JAMA. 2018 May 8;319(18):1914–31.
33. Kilpeläinen TP, Tammela TLJ, Määttänen L, Kujala P, Stenman U-H, Ala-Opas M, et al. False-positive screening results in the Finnish prostate cancer screening trial. Br J Cancer. 2010 Feb;102(3):469–74.
34. Cavazos TB, Witte JS. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. HGG Adv. 2021 Jan 14;2(1):100017.
35. Ankerst DP, Straubinger J, Selig K, Guerrios L, De Hoedt A, Hernandez J, et al. A Contemporary Prostate Biopsy Risk Calculator Based on Multiple Heterogeneous Cohorts. Eur Urol. 2018 Aug 1;74(2):197–203.
36. Kranse R, Roobol M, Schröder FH. A graphical device to represent the outcomes of a logistic regression analysis. The Prostate. 2008;68(15):1674–80.

Figures

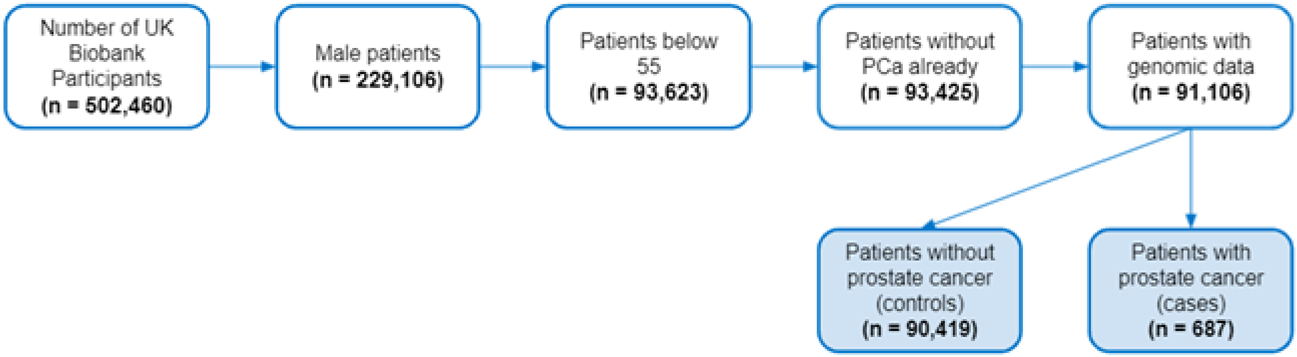
Number of UK Biobank Participants (n = 502,460)
Male patients (n = 229,106)
Patients below 55 (n = 93,623)
Patients without PCa already (n = 93,425)
Patients with genomic data (n = 91,106)
Patients without prostate cancer (controls) (n = 90,419)
Patients with prostate cancer (cases) (n = 687)

Figure 1

Attrition chart for inclusion criteria of UK Biobank participants.

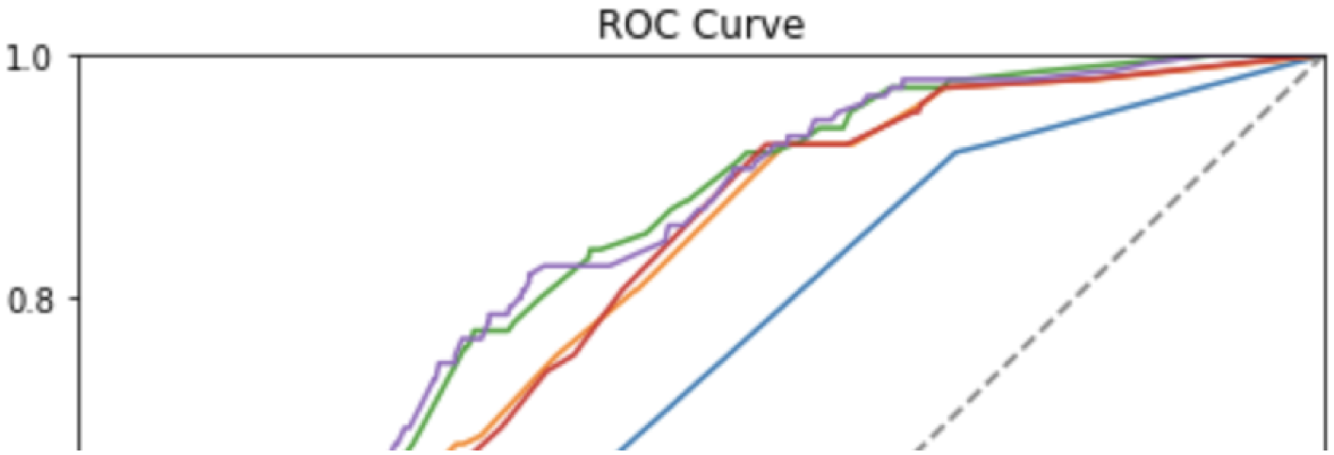

Figure 2

Receiver operating characteristic (ROC) curves of the five machine learning algorithm models for risk prediction of prostate cancer (PCa): PRSs only, Features I, PRSs + Features I, Minimal Features, and PRSs + Minimal Features.

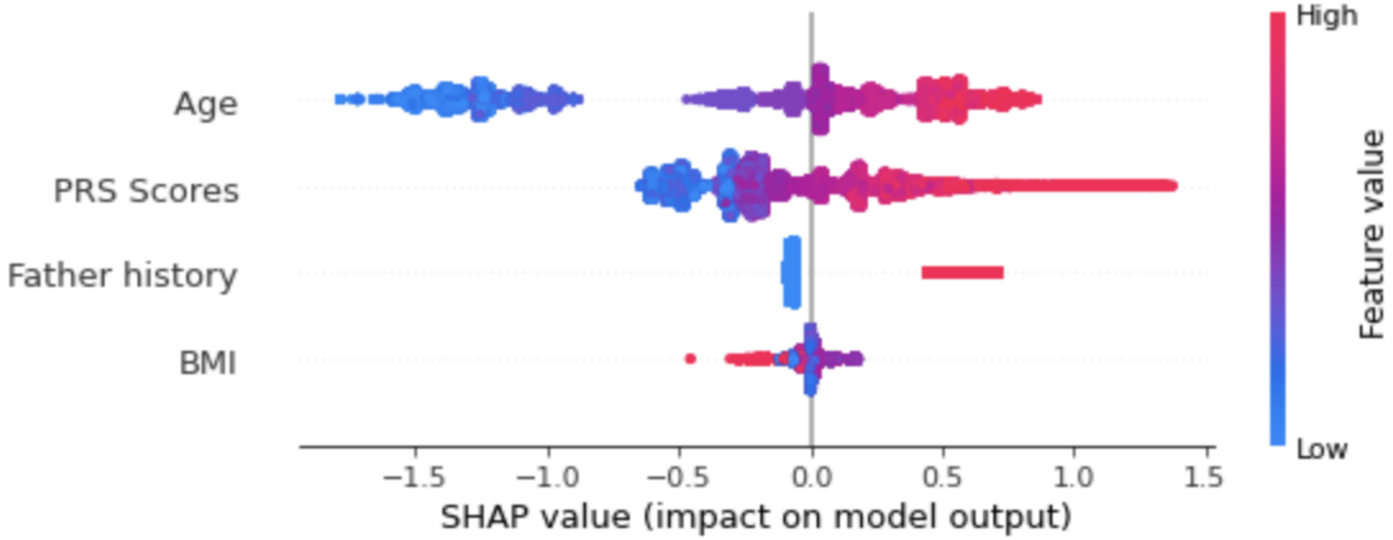

Figure 3

SHapely Additive ExPlanations (SHAP) plot of the PRS + Minimal Features model. Display of the top predictor correlations and distribution of feature importance. Abbreviations: polygenic risk score (PRS); body mass index (BMI).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplemental.docx