Additional SNPs improve risk stratification of a polygenic hazard score for prostate cancer

Roshan A. Karunamuni[1], Minh-Phuong Huynh-Le[2], Chun C. Fan[3], Wesley Thompson[4], Rosalind A. Eeles[5,6], Zsofia Kote-Jarai[5], Kenneth Muir[7,8], Artitaya Lophatananon[7], UKGPCS collaborators*[9], Johanna Schleutker[10,11], Nora Pashayan[12,13], Jyotsna Batra[14,15], APCB BioResource (Australian Prostate Cancer BioResource)*[16,15], Henrik Grönberg[17], Eleanor I. Walsh[18], Emma L. Turner[18], Athene Lane[18,19], Richard M. Martin[18,19,20], David E. Neal[21,22,23], Jenny L. Donovan[24], Freddie C. Hamdy[21,25], Børge G. Nordestgaard[26,27], Catherine M. Tangen[28], Robert J. MacInnis[29,30], Alicja Wolk[31,32], Demetrius Albanes[33], Christopher A. Haiman[34], Ruth C. Travis[35], Janet L. Stanford[36,37], Lorelei A. Mucci[38], Catharine M. L. West[39], Sune F. Nielsen[40,26], Adam S. Kibel[41], Fredrik Wiklund[17], Olivier Cussenot[42,43], Sonja I. Berndt[33], Stella Koutros[33], Karina Dalsgaard Sørensen[44,45], Cezary Cybulski[46], Eli Marie Grindedal[47], Jong Y. Park[48], Sue A. Ingles[49], Christiane Maier[50], Robert J. Hamilton[51,52], Barry S. Rosenstein[53,54], Ana Vega[55,56,57], The IMPACT Study Steering Committee and Collaborators*[58], Manolis Kogevinas[59,60,61,62], Kathryn L. Penney[63], Manuel R. Teixeira[64,65,66], Hermann Brenner[67,68,69], Esther M. John[70], Radka Kaneva[71], Christopher J. Logothetis[72], Susan L. Neuhausen[73], Azad Razack[74], Lisa F. Newcomb[36,75], Canary PASS Investigators*[36,75], Marija Gamulin[76], Nawaid Usmani[77,78], Frank Claessens[79], Manuela Gago-Dominguez[80,81], Paul A. Townsend[82], Monique J. Roobol[83], Wei Zheng[84], The Profile Study Steering Committee*[85], Ian G. Mills[86], Ole A. Andreassen[87], Anders M. Dale[88], Tyler M. Seibert[1,88,89], The PRACTICAL Consortium*[90]

[1]Department of Radiation Medicine and Applied Sciences, University of California San Diego, La Jolla, CA, USA
[2]Radiation Oncology, George Washington University, Washington, DC
[3]Center for Human Development, University of California San Diego, La Jolla, CA, USA
[4]Department of Family Medicine and Public Health, University of California, San Diego, La Jolla, CA, USA
[5]The Institute of Cancer Research, London, SM2 5NG, UK
[6]Royal Marsden NHS Foundation Trust, London, SW3 6JJ, UK
[7]Division of Population Health, Health Services Research and Primary Care, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[8]Warwick Medical School, University of Warwick, Coventry, CV4 7AL, UK
[9]http://www.icr.ac.uk/our-research/research-divisions/division-of-genetics-and-epidemiology/oncogenetics/research-projects/ukgpcs/ukgpcs-collaborators
[10]Institute of Biomedicine, University of Turku, Finland
[11]Department of Medical Genetics, Genomics, Laboratory Division, Turku University Hospital, PO Box 52, 20521 Turku, Finland
[12]Department of Applied Health Research, University College London, London, WC1E 7HB, UK
[13]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Strangeways Laboratory, Worts Causeway, Cambridge, CB1 8RN, UK
[14]Australian Prostate Cancer Research Centre-Qld, Institute of Health and Biomedical

Innovation and School of Biomedical Sciences, Queensland University of Technology, Brisbane QLD 4059, Australia
[15]Translational Research Institute, Brisbane, Queensland 4102, Australia
[16]Australian Prostate Cancer Research Centre-Qld, Queensland University of Technology, Brisbane; Prostate Cancer Research Program, Monash University, Melbourne; Dame Roma Mitchell Cancer Centre, University of Adelaide, Adelaide; Chris O'Brien Lifehouse and
[17]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, SE-171 77 Stockholm, Sweden
[18]Bristol Medical School, Department of Population Health Sciences, University of Bristol, Bristol, United Kingdom
[19]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom
[20]National Institute for Health Research (NIHR) Bristol Biomedical Research Centre, University Hospitals Bristol NHS Foundation Trust and the University of Bristol, Bristol, United Kingdom
[21]Nuffield Department of Surgical Sciences, University of Oxford, Room 6603, Level 6, John Radcliffe Hospital, Headley Way, Headington, Oxford, OX3 9DU, UK
[22]University of Cambridge, Department of Oncology, Box 279, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK
[23]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge, CB2 0RE, UK
[24]School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom
[25]Faculty of Medical Science, University of Oxford, John Radcliffe Hospital, Oxford, United Kingdom
[26]Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark
[27]Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, 2200 Copenhagen, Denmark
[28]SWOG Statistical Center, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[29]Cancer Epidemiology Division, Cancer Council Victoria, 615 St Kilda Road, Melbourne, VIC 3004, Australia
[30]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Grattan Street, Parkville, VIC 3010, Australia
[31]Unit of Cardiovascular and Nutritional Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, SE-171 77 Stockholm, Sweden
[32]Department of Surgical Sciences, Uppsala University, 75185 Uppsala, Sweden
[33]Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, Maryland, 20892, USA
[34]Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, CA 90015, USA
[35]Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, OX3 7LF, UK
[36]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, 98109-1024, USA
[37]Department of Epidemiology, School of Public Health, University of Washington, Seattle,

Washington 98195, USA
[38]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA
[39]Division of Cancer Sciences, University of Manchester, Manchester Academic Health Science Centre, Radiotherapy Related Research, The Christie Hospital NHS Foundation Trust, Manchester, M13 9PL UK
[40]Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, 2200 Copenhagen, Denmark
[41]Division of Urologic Surgery, Brigham and Womens Hospital, 75 Francis Street, Boston, MA 02115, USA
[42]Sorbonne Universite, GRC n°5 , AP-HP, Tenon Hospital, 4 rue de la Chine, F-75020 Paris, France
[43]CeRePP, Tenon Hospital, F-75020 Paris, France.
[44]Department of Molecular Medicine, Aarhus University Hospital, Palle Juul-Jensen Boulevard 99, 8200 Aarhus N, Denmark
[45]Department of Clinical Medicine, Aarhus University, DK-8200 Aarhus N
[46]International Hereditary Cancer Center, Department of Genetics and Pathology, Pomeranian Medical University, 70-115 Szczecin, Poland
[47]Department of Medical Genetics, Oslo University Hospital, 0424 Oslo, Norway
[48]Department of Cancer Epidemiology, Moffitt Cancer Center, 12902 Magnolia Drive, Tampa, FL 33612, USA
[49]Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, CA 90015, USA
[50]Humangenetik Tuebingen, Paul-Ehrlich-Str 23, D-72076 Tuebingen, Germany
[51]Dept. of Surgical Oncology, Princess Margaret Cancer Centre, Toronto ON M5G 2M9, Canada
[52]Dept. of Surgery (Urology), University of Toronto, Canada
[53]Department of Radiation Oncology and Department of Genetics and Genomic Sciences, Box 1236, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, New York, NY 10029, USA
[54]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029-5674 , USA.
[55]Fundación Pública Galega Medicina Xenómica, Santiago de Compostela, 15706, Spain.
[56]Instituto de Investigación Sanitaria de Santiago de Compostela, Santiago De Compostela, 15706, Spain.
[57]Centro de Investigación en Red de Enfermedades Raras (CIBERER), Spain
[58]http://impact.icr.ac.uk
[59]ISGlobal, Barcelona, Spain
[60]IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain
[61]Universitat Pompeu Fabra (UPF), Barcelona, Spain
[62]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
[63]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital/Harvard Medical School, Boston, MA 02184, USA
[64]Department of Genetics, Portuguese Oncology Institute of Porto (IPO-Porto), 4200-072 Porto,

Portugal
[65]Biomedical Sciences Institute (ICBAS), University of Porto, 4050-313 Porto, Portugal
[66]Cancer Genetics Group, IPO-Porto Research Center (CI-IPOP), Portuguese Oncology Institute of Porto (IPO-Porto), 4200-072 Porto, Portugal
[67]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), D-69120, Heidelberg, Germany
[68]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), D-69120 Heidelberg, Germany
[69]Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Im Neuenheimer Feld 460, 69120 Heidelberg, Germany
[70]Departments of Epidemiology & Population Health and of Medicine, Division of Oncology, Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA 94304 USA
[71]Molecular Medicine Center, Department of Medical Chemistry and Biochemistry, Medical University of Sofia, Sofia, 2 Zdrave Str., 1431 Sofia, Bulgaria
[72]The University of Texas M. D. Anderson Cancer Center, Department of Genitourinary Medical Oncology, 1515 Holcombe Blvd., Houston, TX 77030, USA
[73]Department of Population Sciences, Beckman Research Institute of the City of Hope, 1500 East Duarte Road, Duarte, CA 91010, 626-256-HOPE (4673)
[74]Department of Surgery, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia
[75]Department of Urology, University of Washington, 1959 NE Pacific Street, Box 356510, Seattle, WA 98195, USA
[76]Division of Medical Oncology, Urogenital Unit, Department of Oncology, University Hospital Centre Zagreb, University of Zagreb, School of Medicine, 10 000 Zagreb, Croatia
[77]Department of Oncology, Cross Cancer Institute, University of Alberta, 11560 University Avenue, Edmonton, Alberta, Canada T6G 1Z2
[78]Division of Radiation Oncology, Cross Cancer Institute, 11560 University Avenue, Edmonton, Alberta, Canada T6G 1Z2
[79]Molecular Endocrinology Laboratory, Department of Cellular and Molecular Medicine, KU Leuven, BE-3000, Belgium
[80]Genomic Medicine Group, Galician Foundation of Genomic Medicine, Instituto de Investigacion Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario Universitario de Santiago, Servicio Galego de Saúde, SERGAS, 15706, Santiago de Compostela, Spai
[81]University of California San Diego, Moores Cancer Center, Department of Family Medicine and Public Health, University of California San Diego, La Jolla, CA 92093-0012, USA
[82]Division of Cancer Sciences, Manchester Cancer Research Centre, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, NIHR Manchester Biomedical Research Centre, Health Innovation Manchester, Univeristy of Manchester, M13 9WL
[83]Department of Urology, Erasmus University Medical Center, 3015 CE Rotterdam, The Netherlands
[84]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, 2525 West End Avenue, Suite 800, Nashville, TN 37232 USA.
[85]http://www.cancerresearchuk.org/about-cancer/find-a-clinical-trial/a-study-find-out-looking-

gene-changes-would-be-useful-in-screening-for-prostate-cancer-profile-pilot
[86]Center for Cancer Research and Cell Biology, Queen's University of Belfast, Belfast, UK
[87]NORMENT, KG Jebsen Centre, Oslo University Hospital and University of Oslo, Oslo, Norway
[88]Department of Radiology, University of California San Diego, La Jolla, CA, USA
[89]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA
[90]Institute of Cancer Research, Sutton, SW7 3RPm UK

* Additional consortia memebers, associated funding, and contact information are provided in the Supplementary Data Appendix 1,2, and 3.

Corresponding Authors:

Roshan Karunamuni, PhD
University of California, San Diego
Department of Radiation Medicine and Applied Sciences
3960 Health Sciences Dr, Mail Code 0865
La Jolla, CA 92093
rakarunamuni@health.ucsd.edu
ORCID: 0000-0001-8723-8123

Tyler M. Seibert, MD, PhD
University of California, San Diego
Department of Radiation Medicine and Applied Sciences
9500 Gilman Dr, Mail Code 0861
La Jolla, CA 92093
tseibert@ucsd.edu
ORCID: 0000-0002-4089-7399

Running Title

Adding SNPs improves prostate cancer polygenic score

Funding

This study was funded in part by a grant from the United States National Institute of Health/National Institute of Biomedical Imaging and Bioengineering (#K08EB026503), the

University of California Cancer Research Coordinating Committee (#C21CR2060), the Research Council of Norway (#223273), KG Jebsen Stiftelsen, and South East Norway Health Authority.

Abstract

Background: Polygenic hazard scores (PHS) can identify individuals with increased risk of prostate cancer. We estimated the benefit of additional SNPs on performance of a previously validated PHS (PHS46).

Materials and Method: 180 SNPs, shown to be previously associated with prostate cancer, were used to develop a PHS model in men with European ancestry. A machine-learning approach, LASSO-regularized Cox regression, was used to select SNPs and to estimate their coefficients in the training set (75,596 men). Performance of the resulting model was evaluated in the testing/validation set (6,411 men) with two metrics: (1) hazard ratios (HRs) and (2) positive predictive value (PPV) of prostate-specific antigen (PSA) testing. HRs were estimated between individuals with PHS in the top 5% to those in the middle 40% (HR95/50), top 20% to bottom 20% (HR80/20), and bottom 20% to middle 40% (HR20/50). PPV was calculated for the top 20% (PPV80) and top 5% (PPV95) of PHS as the fraction of individuals with elevated PSA that were diagnosed with clinically significant prostate cancer on biopsy.

Results: 166 SNPs had non-zero coefficients in the Cox model (PHS166). All HR metrics showed significant improvements for PHS166 compared to PHS46: HR95/50 increased from 3.72 to 5.09, HR80/20 increased from 6.12 to 9.45, and HR20/50 decreased from 0.41 to 0.34. By contrast, no significant differences were observed in PPV of PSA testing for clinically significant prostate cancer.

Conclusion: Incorporating 120 additional SNPs (PHS166 vs PHS46) significantly improved HRs for prostate cancer, while PPV of PSA testing remained the same.

Introduction

Optimal prostate cancer screening strategies seek to strike a balance between identifying clinically significant and potentially lethal cases that require treatment, while minimizing overdiagnosis of indolent, lower-risk cases that do not need radical treatment[1–3]. Genetic risk models have emerged as potentially useful tools that identify individuals with greater risk for being diagnosied with prostate cancer[4,5], and so help inform if and when to initiate screening for an individual. A subset of these models called polygenic hazard scores (PHS) seeks to directly identify associations between common genetic variants and the age of diagnosis of prostate cancer by utilizing the framework of time-to-event analyses[1,6].

We have previously reported on a PHS model for prostate cancer, PHS46, that demonstrated excellent performance in an independent test set of men from varied genetic ancestries[6]. The model incorporates genetic data of 46 unique single nucleotide polymorphisms (SNPs), and was identified through a systematic search of European men genotyped on the iCOGS chipset (Illumina, San Diego, CA). With an ever-increasing list of loci associated with prostate cancer in the literature[7–9], we sought to determine what effect, if any, the incorporation of additional SNPs would have on the performance of PHS46.

To this end, we employed a machine-learning approach, LASSO-regularized Cox regression,[10,11] to select SNPs from a list that included the 46 used in PHS46, as well as over 100 SNPs identified in previous analyses as having genome-wide significance for association with prostate cancer[7]. LASSO-regularized regression is an established variable selection technique in datasets with a large number of predictors and has been previously implemented as a SNP selection tool for a breast cancer polygenic risk score[12]. Performance metrics describing statistical model goodness-of-fit and clinically actionable screening utility of the LASSO-regularized PHS model for prostate cancer were compared with those achieved with PHS46 to determine the potential benefit of incorporating additional SNPs in polygenic hazard models.

Material and Methods

Study dataset

We obtained genotype and phenotype data from the PRACTICAL[13] consortium for this analysis. Genotyping was performed previously on either OncoArray[13] or iCOGS[9] chips, and these data were previously imputed using the 1000 Genomes reference panel[14]. Missing SNP calls were replaced with the mean of the genotyped data for that SNP in the training set[1,15]. In total, data from 82,007 men with European genetic ancestry (Supplementary Table 1)[13,16] were available for this analysis. A testing set consisting of 6,411 men (4,828 controls and 1,583 cases) enrolled in the ProtecT clinical trial was set aside for estimating the performance of the final PHS models. The data from ProtecT were chosen as the testing set because they are well characterized and were previously used for validation of PHS46[1], allowing us to directly benchmark the performance of the updated model against previous iterations. The ProtecT trial also included biopsies of participants with elevated prostate-specific antigen (PSA) level, which permits analysis of the positive predictive value of the current clinical standard for screening, PSA testing. The remaining 75,596 individuals (25,127 controls and 50,469 cases) were used for training of the model. This first analysis was limited to men of European descent because of much greater data availability in that population, but our previous work has shown that development in Europeans can inform careful future work to assess and improve performance in other ancestries[17].

Model development using LASSO regularization

A list of published SNPs previously identified[1,7] to be associated with prostate cancer was compiled. In total, 180 unique SNPs were considered for estimation within the PHS model framework. An initial screening was conducted to identify pairs of SNPs that were highly correlated ($R^2 > 0.95$). For each pair of highly correlated SNPs, a univariable Cox proportional hazards model using age of diagnosis of prostate cancer as the time to event was calculated for

each SNP in the pair, and the one with the larger p-value was discarded. The remaining SNPs were included as candidates for the new PHS model. The R (v.4.0.1) package 'glmnet' was used to estimate a LASSO-regularized Cox-proportional hazards model[10,11] using age of diagnosis of prostate cancer as the time to event. The genetic data of candidate SNPs and first four European ancestry principal components were included as predictors. Controls were censored at age of last follow-up. The hyper-parameter of the LASSO-regularized model, lambda, was selected using 10-fold cross-validation[10,11]. The final form of the LASSO model was estimated at the value of lambda that minimized the mean cross-validated error.

Characterization of LASSO-regularized PHS model

The PHS score for each of the individuals in the training and testing set was estimated as the weighted sum of the genetic counts of each of the SNPs in the PHS model, using the LASSO model coefficients as weights. Distributions of the new PHS score were compared qualitatively between training and testing groups to confirm that the model was appropriately calibrated for use in the testing set.

We also sought to assess how the LASSO-regularized PHS score compared to family history in explaining the variation in age at diagnosis of prostate cancer. A multivariable Cox proportional hazards model was estimated using the age at diagnosis of any prostate cancer as the time to event, and the PHS score and family history as predictors in both training and testing sets, separately. The family history variable was coded as a binary variable: "None" or "One or more affected first-degree relatives". Observations with missing family history values were removed from the analysis. The explained relative risk[18] (ERR) of each of the covariables as well as the full model were estimated using the "clinfun" software package in R, and provided a quantifiable measure for the importance of each variable in the model. Empirical confidence intervals for ERR were estimated using 1000 bootstrapped iterations.

Performance comparison between PHS46 and LASSO-regularized PHS

Performance in the testing set was assessed using hazard ratios (HRs) and positive predictive value (PPV), as described below. In each case, performance metrics were generated for the newly developed LASSO PHS model and for PHS46. Model coefficients for PHS46 were obtained from the literature[17]. For each performance metric, one thousand bootstrap samples of the testing set were used to generate empirical 95% confidence intervals for LASSO PHS and for PHS46. In addition, bootstrapped 95% confidence intervals were generated for the percentage change of each performance metric between the two models, using PHS46 as the reference. Percent changes were deemed statistically significant if the bootstrapped 95% confidence interval did not include 0.

HR performance

Calibration Cox proportional hazards models were fit to the bootstrapped testing data using the PHS score as the sole predictor and the age-of-diagnosis of prostate cancer as the dependent variable. The model coefficient of this Cox regression model is referred to as the calibration factor. Next, the hazard ratio between two PHS groups, such as those in the top 5% to the middle 40% (HR95/50), is estimated as the exponential of the product of the calibration factor and the difference in mean PHS scores of each group. Hazard ratios between the top 20% to the bottom 20% (HR80/20) and the bottom 20% to the middle 40% (HR20/50) were similarly calculated. The PHS cutoffs used to define these groups were determined from the distribution of PHS in the training set controls under 70 years of age[1,15].

A similar strategy was used to estimate the HR performance for clinically significant prostate cancer. The criteria for clinical significance were any of: Gleason score >=7, stage T3-T4, PSA concentration >= 10ng/mL, pelvic lymph nodal metastasis, or distant metastasis[19]. In this analysis, controls and low-risk (i.e., not clinically significant) cancers were censored at age of last follow-up and age of diagnosis, respectively. HRs are reported after sample-weight

correction[1,17,20] using the total number of cases and controls in the ProtecT trial to generate weighting factors.

Sample-weight corrected HR values were generated using the age at diagnosis of non-clinically significant prostate cancer. Individuals with clinically significant prostate cancer were removed from this secondary analysis.

PPV performance

One indicator of clinical utility of a risk-stratification approach like PHS is whether it can be used to improve the PPV of the standard clinical screening test, prostate-specific antigen (PSA). As a population-based screening study, ProtecT provides biopsy results of both cases and controls with a positive PSA result (i.e., ≥3 ng/mL). PPV performance of each model was estimated by randomly sampling individuals within the testing set with positive PSA results, while maintaining the case to control ratio of the ProtecT study (1:2). PPV is calculated as the fraction of positive PSA individuals in the top 20% (PPV80) or top 5% (PPV95) of PHS scores that had clinically significant prostate cancer.

Cumulative incidence curves for LASSO-PHS in United Kingdom

To illustrate the utility of the LASSO PHS model in informing prostate cancer screening, cumulative incidence curves for various PHS risk groups were estimated, as described previously[21]. The age-specific general cumulative incidence curve for prostate cancer was estimated for the United Kingdom population, aged 40 to 70, using data from Cancer Research UK 2015-2017[22]. The proportion of clinically significant and non-clinically significant prostate cancer at each age was estimated using data from the Cluster Randomized Trial of PSA Testing for Prostate Cancer (CAP) trial[23]. Disease-specific cumulative incidence curves for clinically significant and non-clinically-significant prostate cancer were estimated by multiplying the general cumulative incidence curve by their respective proportions. The risk-adjusted incidence

curves for individuals in the upper 5th percentile and upper 20th percentile were estimated by multiplying the disease-specific cumulative incidence curves by the mean value of HR95/50 and HR80/50 in the testing set, respectively. Hazard ratios were obtained using the age of diagnosis of clinically significant prostate cancer as the time-to-event and after sample-weight correction.

Results

SNP screening and PHS model training

Of the 180 SNPs originally considered for this study, 6 SNPs were discarded in the initial screening process of removing highly correlated SNPs. Of the 174 remaining candidate SNPs (Supplementary Table 2), 166 had non-zero LASSO model coefficients and were selected for the final PHS model (PHS166).

The majority of the 166 variants (97, 53%) used in PHS166 were classified as intron variants (Supplementary Table 3). Of the genes associated with variants from PHS166, HNF1B on chromosome 17 was associated with the greatest number of variants (4). Additional genes that were associated with multiple variants included ITGA6(x2), LINC00506(x2), PDLIM5(x2), TERT(x2), CTD-2194D22.4(x2), RGS17(x2), LOC105375751(x2), and CASC8(x3). Two of the SNPs used in PHS166 (rs721048 and rs10993994) were designated as 'pathogenic' by ClinVar[24] and associated with hereditary prostate cancer.

PHS166 model characterization

Distributions of PHS166 score were visually consistent between training and testing sets (Supplementary Figure 1). The 20th, 30th, 70th, 80th, and 98th percentiles of the reference PHS risk scores (controls in training set) were estimated as -0.411, -0.307, 0.048, 0.154, and 0.557, respectively.

PHS166 contributed roughly 80 to 90 percent of the total explained relative risk (Supplementary Table 4) of a Cox proportional hazards model containing both family history and PHS166. Family history was not found to be statistically significantly associated with age at diagnosis of prostate cancer in the testing set[1].

Performance comparison – PHS46 vs. PHS166

All PHS166 HR-based performance metrics showed statistically significant improvements compared to PHS46 (Table 1), for both any and clinically significant prostate cancer. The mean HR95/50 and HR80/20 values for PHS166 were roughly 36 to 55% greater than those for PHS46. For example, HR80/20 for clinically significant prostate cancer increased from 6.12 to 9.45. Similarly, HR20/50 for PHS166 was, on average, 18% lower than that for PHS46. Similar trends were observed for non-clinically significant prostate cancer (Supplementary Table 5). No significant differences between models were observed in either of the PPV-based performance metrics (Table 2). Among individuals in the top 20% of risk scores with a positive PSA test, the estimated mean PPV for clinically significant prostate cancer was roughly 0.19 irrespective of the model used – indicating approximately 19% of positive PSA tests in this risk group yielded a diagnosis of clinically significant prostate cancer. By comparison, approximately 13% of all positive PSA tests resulted in a diagnosis of clinically significant prostate cancer.

Cumulative incidence curves for PHS166 in United Kingdom

Cumulative incidence curves for clinically significant and non-clinically significant prostate cancer for the upper 5th percentile (>95th percentile) and upper 20th percentile (>80th percentile) of PHS166 scores in the United Kingdom demonstrated expected stratification of prostate cancer risk (Figure 1).

Discussion

Using a machine-learning, LASSO-regularized Cox framework, we identified 166 SNPs to be included in a polygenic hazard model (PHS166) for association with age of diagnosis of prostate cancer in men of European genetic ancestry. Variants used in PHS166 were associated with several genes, including those encoding for hepatocyte nuclear factor-1 beta (HNF1B), cancer susceptibility 8 (CASC8), and telomerase (TERT). PHS166 also explained a much larger percentage of the total explained relative risk compared to family history, suggesting that the former is important for stratifying patients' risk. When compared to the original PHS, consisting of 46 SNPs, PHS166 demonstrated substantially improved HR performance. For example, the HR for clinically significant prostate cancer comparing the upper and lower quintiles of genetic risk increased by 56% when using PHS166. No significant improvements were found in the PPV of PSA testing when using PHS to stratify risk.

Increased separation in hazard rates between PHS risk groups may allow for more nuance in clinical decision making in certain scenarios. Accurate identification of low, intermediate, and high PHS risk groups in prostate cancer may help in decisions of when (or if) to initiate screening as well as possibly improving the interpretation of the disease screens[25]. Targeting screening to men in the upper percentiles of polygenic risk as opposed to those in the lowest risk group may reduce the proportion of overdiagnosed indolent cancers from 43% to 19%[26,27]. Risk stratification achieved here by PHS166 is similar or better than commonly used clinical tools for diseases such as breast cancer, diabetes, and cardiovascular disease[25,28–30]. Clinically meaningful risk stratification is illustrated by the estimated cumulative incidence curves in Figure 1. This effect is particularly pronounced for clinically significant disease because of the increased proportion of clinically significant cases observed at older ages[2,21,23].

The lack of improvement in PPV in this study may suggest a "performance plateau" when using PHS to define broad risk categories for certain clinical applications. A similar effect has been previously described for prostate cancer polygenic models, in the context of using risk

scores to discriminate prostate biopsy outcomes[31]. Some of the precision in a score may also be diluted in broad clinical applications. The PPV analysis here is applied to participants in the ProtecT trial, which enrolled men aged 50 to 69 years, and screening in the trial was offered irrespective of underlying genetic risk[2]. Further investigation is needed to learn whether timing screening according to genetic risk might better leverage the superior HR performance of PHS166 risk score to improve the PPV of PSA testing.

LASSO frameworks have been used to identify SNPs for polygenic risk scores of several phenotypes, including fracture risk[32], type 2 diabetes[33], and breast cancer[12]. In this work, we have extended the application of LASSO to select SNPs in a polygenic hazard model of prostate cancer from a list of candidates previously identified through logistic and time-to-event analysis. Simulation studies[11] have suggested that LASSO provides more robust estimates than stepwise selection in cases with both a few large effects, as well as many small effects. As new prostate cancer associated variants are discovered, this framework can be easily implemented to develop updated polygenic hazard models.

One limitation of PHS166 is that it was entirely developed and tested in European men. However, a well-vetted, well-tested PHS model for men of European genetic ancestry can be used as a starting block for developing models for other genetic ancestries, where large-scale databases are often more scarce, as has been shown for PHS46[17,34]. Furthermore, some of the SNPs selected for incorporation into PHS166 were originally discovered in analyses that included men from the ProtecT testing set. Therefore, the improvements in HRs observed for PHS166 may be overestimated. However, this bias is likely small, given that the testing set was only a small fraction (less than 5%) of the data used in prior discovery analyses, and the ProtecT data were not used to calculate SNP weights in PHS166. The LASSO-regularized Cox framework was also used to minimize any potential for over-fitting[35] by introducing penalties for large effect sizes. In addition, this study uses age of diagnosis as the time-to-event variable, and

any preceding period of undiagnosed disease is unknown. Hypothetical perfect measurement of age of onset would likely further improve performance of the PHS model.

In conclusion, we applied a machine-learning, LASSO-regularized Cox regression framework to develop a larger PHS that includes 166 previously discovered SNPs. When comparing the performance of PHS166 to the original model, PHS46, we found that incorporating 120 more SNPs significantly improved HRs for clinically significant prostate cancer. However, incorporating more SNPs did not improve on the ability of PHS46 to inform the PPV of PSA testing in the ProtecT dataset, perhaps illustrating a plateau effect and/or dilution of risk stratification in a broad clinical application.

Ethics Statement

All contributing studies were approved by the relevant ethics committees and performed in accordance with the Declaration of Helsinki; written informed consent was obtained from the study participants. The present analyses used de-identified data from the PRACTICAL consortium and have been approved by the review board at the corresponding authors' institution.

Conflict of Interest:

All authors declare no personal or financial conflicts of interest for the submitted work except as follows. CCF is a scientific consultant for CorTechs Labs, Inc. RE reports honorarium as a speaker for GU-ASCO meeting in San Francisco Jan 2016, support from Janssen, and honorarium as speaker for RMH-FR meeting Nov 2017. She reports honorarium as a speaker at the University of Chicago invited talk May 2018, and an educational honorarium by Bayer & Ipsen to attend GU Connect “Treatment sequencing for mCRPC patients within the changing landscape of mHSPC” at ESMO Barcelona, Sep 2019. She reports member of external Expert Committee on the Prostate Dx Advisory Panel. OAA received speaker’s honorarium from Lundbeck, and is a consultant for Healthlytix. AMD reports that he was a founder and holds equity in CorTechs Labs Inc., and serves on its Scientific Advisory Board. He is a member of the Scientific Advisory Board of Human Longevity, Inc., and the Mohn Medical Imaging and Visualization Centre. He received funding through research grants from GE Healthcare to UCSD. The terms of these arrangements have been reviewed by and approved by UCSD in accordance with its conflict of interest policies. TMS reports honoraria, outside of the present work, from: University of Rochester, Varian Medical Systems, Multimodal Imaging Servcies Corporation; and WebMD. He reports research funding from NIH/NBIB, U.S. Department of Defense, Radiological Society of North America, American Society for Radiation Oncology, and Varian Medical Systems.

Data Availability Statement

The data used in this work were obtained from the Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL) consortium. Readers who are interested in accessing the data must first submit a proposal to the Data Access Committee. If the reader is not a member of the consortium, their concept form must be sponsored by a principal investigator (PI) of one of the PRACTICAL consortium member studies. If approved by the Data Access Committee, PIs within the consortium, each of whom retains ownership of their data submitted to the consortium, can then choose to participate in the specific proposal. In addition, portions of the data are available for request from dbGaP (database of Genotypes and Phenotypes) which is maintained by the National Center for Biotechnology Information (NCBI): https://www.ncbi.nlm.nih.gov/gap/?term=Icogs+prostatehttps://www.ncbi.nlm.nih.gov/gap/?term=Icogs+prostate.

Anyone can apply to join the consortium. The eligibility requirements are listed here: http://practical.icr.ac.uk/blog/?page_id=9. Joining the consortium would not guarantee access, as a proposal for access would still be submitted to the Data Access Committee, but there would be no need for a separate member sponsor. Readers may find information about application by using the contact information below:

Rosalind Eeles
Principal Investigator for PRACTICAL
Professor of Oncogenetics
Institute of Cancer Research (ICR)
Sutton, UK
Email: PRACTICAL@icr.ac.uk
URL: http://practical.icr.ac.uk
Tel: ++44 (0)20 8722 4094

References

1 Seibert TM, Fan CC, Wang Y, Zuber V, Karunamuni R, Parsons JK *et al.* Polygenic hazard score to guide screening for aggressive prostate cancer: Development and validation in large scale cohorts. *BMJ* 2018; **360**: 1–7.

2 Huynh-Le MP, Myklebust TÅ, Feng CH, Karunamuni R, Johannesen TB, Dale AM *et al.* Age dependence of modern clinical risk groups for localized prostate cancer—A population-based study. *Cancer* 2020; **126**: 1691–1699.

3 Pashayan N, Duffy SW, Chowdhury S, Dent T, Burton H, Neal DE *et al.* Polygenic susceptibility to prostate and breast cancer: Implications for personalised screening. *Br J Cancer* 2011; **104**: 1656–1663.

4 Witte JS. Personalized prostate cancer screening: Improving PSA tests with genomic information. *Sci Transl Med* 2010; **2**: 1–5.

5 Chen H, Liu X, Brendler CB, Ankerst DP, Leach RJ, Goodman PJ *et al.* Adding genetic risk score to family history identifies twice as many high-risk men for prostate cancer: Results from the prostate cancer prevention trial. *Prostate* 2016; **76**: 1120–1129.

6 Huynh-Le M-P, Fan CC, Karunamuni R, Thompson WK, Martinez ME, Eeles RA *et al.* Polygenic hazard score is associated with prostate cancer in multi-ethnic populations. *medRxiv* 2020; : 1–34.

7 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; **50**: 928–936.

8 Kote-Jarai Z, Easton DF, Stanford JL, Ostrander EA, Schleutker J, Ingles SA *et al.* Multiple novel prostate cancer predisposition loci confirmed by an international study: The PRACTICAL consortium. *Cancer Epidemiol Biomarkers Prev* 2008; **17**: 2052–2061.

9 Eeles RA, Olama AA Al, Benlloch S, Saunders EJ, Leongamornlert DA, Tymrakiewicz M *et al.* Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom

genotyping array. *Nat Genet* 2013; **45**: 385–391.

10 Tibshiranit BR. Regression Shrinkage and Selection via the Lasso. *J R Stat Soc B* 1996; : 267–288.

11 Tibshirani R. The lasso method for variable selection in the cox model. *Stat Med* 1997; **16**: 385–395.

12 Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet* 2019; **104**: 21–34.

13 Amos CI, Dennis J, Wang Z, Byun J, Schumacher FR, Gayther SA *et al.* The oncoarray consortium: A network for understanding the genetic architecture of common cancers. *Cancer Epidemiol Biomarkers Prev* 2017; **26**: 126–135.

14 Eeles R. Prostate cancer genome-wide association study from 89,000 men using the OncoArray chip to identify novel prostate cancer susceptibility loci. *J Clin Oncol* 2016; **34**: 1525.

15 Karunamuni RA, Huynh-Le MP, Fan CC, Eeles RA, Easton DF, Kote-Jarai ZsS *et al.* The effect of sample size on polygenic hazard models for prostate cancer. *Eur J Hum Genet* 2020. doi:10.1038/s41431-020-0664-2.

16 Li Y, Byun J, Cai G, Xiao X, Han Y, Cornelis O *et al.* FastPop: A rapid principal component derived method to infer intercontinental ancestry using genetic data. *BMC Bioinformatics* 2016; **17**: 1–8.

17 Huynh-Le M-P, Chieh Fan C, Karunamuni R, Martinez ME, Eeles RA, Kote-Jarai Z *et al.* Polygenic hazard score is associated with prostate cancer in multi-ethnic populations. *medRxiv* 2019. doi:https://doi.org/10.1101/19012237.

18 Heller G. A measure of explained risk in the proportional hazards model. *Biostatistics* 2012; **13**: 315–325.

19 NCCN Clinical Practice Guidelines in Oncology. Prostate Cancer. Version 1.2019. .

20 Therneau TM, Li H. Computing the Cox Model for Case Cohort Designs. *Lifetime Data Anal* 1999; **5**: 99–112.

21 Huynh-Le M-P, Fan CC, Karunamuni R, Walsh EI, Turner EL, Lane JA *et al.* A genetic risk score to personalize prostate cancer screening, applied to population data. *Cancer Epidemiol Biomarkers Prev* 2020; : cebp.1527.2019.

22 Prostate cancer incidence statistics | Cancer Research UK. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer/incidence#heading-One (accessed 23 Jul2020).

23 Martin RM, Donovan JL, Turner EL, Metcalfe C, Young GJ, Walsh EI *et al.* Effect of a low-intensity PSA-based screening intervention on prostate cancer mortality: The CAP randomized clinical trial. *JAMA - J Am Med Assoc* 2018; **319**: 883–895.

24 Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM *et al.* ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 2014; **42**: 980–985.

25 Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* 2018; **19**: 581–590.

26 Pashayan N, Duffy SW, Neal DE, Hamdy FC, Donovan JL, Martin RM *et al.* Implications of polygenic risk-stratified screening for prostate cancer on overdiagnosis. *Genet Med* 2015; **17**: 789–795.

27 Pashayan N, Pharoah PDP, Schleutker J, Talala K, Tammela TLJ, Määttänen L *et al.* Reducing overdiagnosis by polygenic risk-stratified screening: Findings from the Finnish section of the ERSPC. *Br J Cancer* 2015; **113**: 1086–1093.

28 Wang TJ, Larson MG, Levy D, Benjamin EJ, Leip EP, Omland T *et al.* Plasma Natriuretic Peptide Levels and the Risk of Cardiovascular Events and Death. *N Engl J Med* 2004; **350**: 655–663.

29 Yang X, Leslie G, Gentry-Maharaj A, Ryan A, Intermaggio M, Lee A *et al.* Evaluation of

polygenic risk scores for ovarian cancer risk prediction in a prospective cohort study. *J Med Genet* 2018; **55**: 546–554.

30 Yeh HC, Duncan BB, Schmidt MI, Wang NY, Brancati FL. Smoking, smoking cessation, and risk for type 2 diabetes mellitus: A cohort study. *Ann Intern Med* 2010; **152**: 10–17.

31 Ren S, Xu J, Zhou T, Jiang H, Chen H, Liu F *et al.* Plateau effect of prostate cancer risk-associated SNPs in discriminating prostate biopsy outcomes. *Prostate* 2013; **73**: 1824–1835.

32 Forgetta V, Keller-baruch J, Forest M, Durand A, Bhatnagar S, Kemp JP *et al.* Development of a polygenic risk score to improve screening for fracture risk : A genetic risk prediction study. *PLoS Med* 2020; : 1–19.

33 Chen T-H, Chatterjee N, Landi MT, Shi J. A penalized regression framework for building polygenic risk models based on summary statistics from genome-wide association studies and incorporating external information. *J Am Stat Assoc* 2020; **1459**: 1–19.

34 Karunamuni R, Huynh-Le M-P, Fan CC, Thompson W, Eeles RA, Kote-Jarai Z *et al.* African-specific improvement of a polygenic hazard score for age at diagnosis of prostate cancer. *medRxiv* 2020; : 1–32.

35 McNeish DM. Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences. *Multivariate Behav Res* 2015; **50**: 471–484.

Figure Legends

Figure 1. Cumulative incidence curves for PHS166. Risk-adjusted cumulative incidence curves for the upper 5th percentile (>95th percentile) and upper 20th percentile (>80th percentile) of PHS166 scores for clinically significant and non-clinically-significant prostate cancer. Reference curves representing the population average cumulative incidence (i.e., unadjusted for genetic risk).

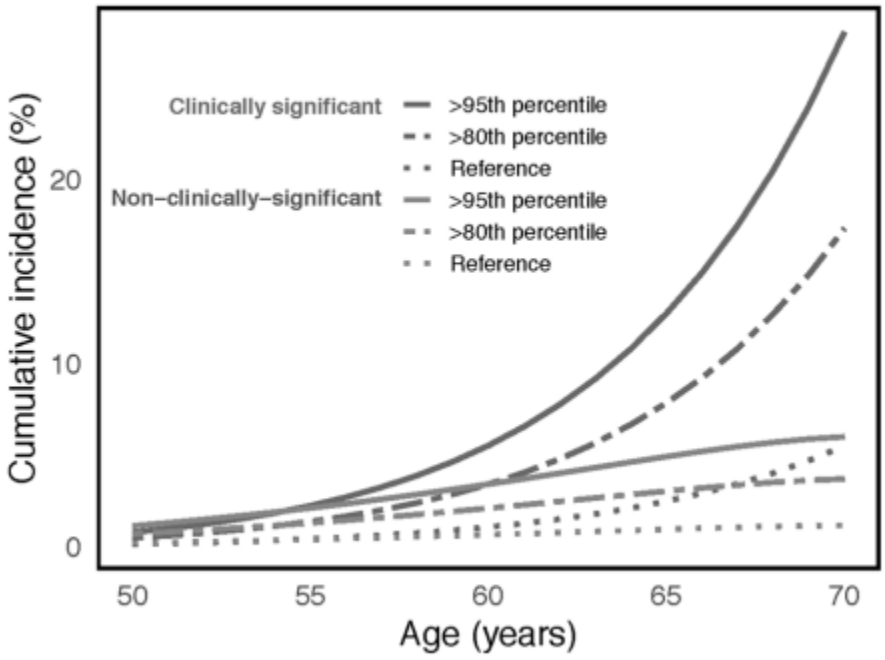

Table 1. HR performance in testing set. Sample-weight-corrected hazard ratios are estimated for PHS166 and PHS46 in the testing set, using age-of-onset of any or clinically significant prostate cancer. The percent change for each metric is calculated using the value of PHS46 as the reference. Mean values and 95% confidence intervals are reported.

Type of cancer	HR	PHS46	PHS166	Change (%)
Any	HR95/50	3.29 [2.73,3.77]	4.45 [3.68,5.06]	36 [18,53]
	HR80/20	5.15 [3.92,6.18]	7.85 [6.04,9.33]	53 [25,78]
	HR20/50	0.44 [0.40,0.49]	0.37 [0.33,0.40]	-18 [-25,-10]
Clinically Significant	HR95/50	3.72 [2.89,4.43]	5.09 [3.84,6.05]	37 [13,59]
	HR80/20	6.12 [4.18,7.67]	9.45 [6.17,11.79]	55 [17,88]
	HR20/50	0.41 [0.35,0.47]	0.34 [0.29,0.39]	-18 [-28,-9]

Table 2. PPV performance in testing set. Positive predictive value (PPV) of PSA testing for clinically significant prostate cancer using top 5% (PPV95) and top 20% (PPV80) cutoffs of PHS166 and PHS46 risk scores. The percent change for each metric is calculated using the value of PHS46 as the reference.

PPV	PHS46	PHS166	Change (%)
PPV95	0.227 [0.159,0.292]	0.239 [0.171,0.305]	6.3 [-25.5,32.1]
PPV80	0.192 [0.155,0.231]	0.187 [0.150,0.222]	-2.8 [-16.3,9.9]