The effect of sample size on polygenic hazard models for prostate cancer

Roshan A. Karunamuni[1], Minh-Phuong Huynh-Le[1], Chun C. Fan[2], Rosalind A. Eeles[3,4], Douglas F. Easton[5], ZSofia Kote-Jarai[3], Ali Amin Al Olama[5,6], Sara Benlloch Garcia[5], Kenneth Muir[7,8], Henrik Gronberg[9], Fredrik Wiklund[9], Markus Aly[9,10,11], Johanna Schleutker[12,13], Csilla Sipeky[12], Teuvo LJ Tammela[14,15], Børge G. Nordestgaard[16,17], Tim J. Key[18], Ruth C. Travis[18], David E. Neal[19,20,21], Jenny L. Donovan[22], Freddie C. Hamdy[23,24], Paul Pharoah[25], Nora Pashayan[26,25,27], Kay-Tee Khaw[28], Stephen N. Thibodeau[29], Shannon K. McDonnell[30], Daniel J. Schaid[30], Christiane Maier[31], Walther Vogel[32], Manuel Luedeke[31], Kathleen Herkommer[33], Adam S. Kibel[34], Cezary Cybulski[35], Dominika Wokolorczyk[35], Wojciech Kluzniak[35], Lisa Cannon-Albright[36,37], Hermann Brenner[38,39,40], Ben Schöttker[41,42], Bernd Holleczek[43,44], Jong Y. Park[45], Thomas A. Sellers[45], Hui-Yi Lin[46], Chavdar Slavov[47], Radka Kaneva[48], Vanio Mitev[48], Jyotsna Batra[49,50], Judith A. Clements[51,52], Amanda Spurdle[53], Australian Prostate Cancer BioResource (APCB)[51], Manuel R. Teixeira[54,55], Paula Paulo[54,56], Sofia Maia[54,56], Hardev Pandha[57], Agnieszka Michael[57], Ian G. Mills[58,59], Ole A. Andreassen[60], Anders M. Dale[61,62,63], Tyler M. Seibert[1], The PRACTICAL Consortium^

[1]Department of Radiation Medicine and Applied Sciences, University of California San Diego, La Jolla, CA, USA
[2]Healthlytix, 4747 Executive Dr. Suite 820, San Diego, CA, USA
[3]The Institute of Cancer Research, London, SM2 5NG, UK
[4]Royal Marsden NHS Foundation Trust, London, SW3 6JJ, UK
[5]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Cambridge CB1 8RN, UK
[6]University of Cambridge, Department of Clinical Neurosciences, Stroke Research Group, R3, Box 83, Cambridge Biomedical Campus, Cambridge CB2

0QQ, 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, UK

[9]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, SE-171 77 Stockholm, Sweden

[10]Department of Molecular Medicine and Surgery, Karolinska Institute, SE-171 77 Stockholm, Sweden

[11]Department of Urology, Karolinska University Hospital, Stockholm, Sweden

[12]Institute of Biomedicine, Kiinamyllynkatu 10, FI-20014 University of Turku, Finland

[13]Department of Medical Genetics, Genomics, Laboratory Division, Turku University Hospital, PO Box 52, 20521 Turku, Finland

[14]Faculty of Medicine and Health Technology, Prostate Cancer Research Center, FI-33014 Tampere University, Finland

[15]Department of Urology, Tampere University Hospital, Tampere, Finland

[16]Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark

[17]Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, 2200 Copenhagen, Denmark

[18]Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, OX3 7LF, UK

[19]Nuffield Department of Surgical Sciences, University of Oxford, Room 6603, Level 6, John Radcliffe Hospital, Headley Way, Headington, Oxford, OX3 9DU, UK

[20]University of Cambridge, Department of Oncology, Box 279, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK

[21]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge UK

[22]School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol, BS8 2PS, UK

[23]Nuffield Department of Surgical Sciences, University of Oxford, Oxford, OX1 2JD, UK

[24]Faculty of Medical Science, University of Oxford, John Radcliffe Hospital, Oxford, UK

[25]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Strangeways Laboratory, Worts Causeway, Cambridge, CB1 8RN, UK

[26]University College London, Department of Applied Health Research, London,UK

[27]Department of Applied Health Research, University College London, London, WC1E 7HB, UK

[28]Clinical Gerontology Unit, University of Cambridge, Cambridge, CB2 2QQ, UK

[29]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA

[30]Division of Biomedical Statistics & Informatics, Mayo Clinic, Rochester, MN

55905, USA
[31]Humangenetik Tuebingen, Paul-Ehrlich-Str 23, D-72076 Tuebingen
[32]Institute for Human Genetics, University Hospital Ulm, 89075 Ulm, Germany
[33]Technical University of Munich, School of Medicine, Klinikum rechts der Isar, Department of Urology
[34]Division of Urologic Surgery, Brigham and Womens Hospital, 75 Francis Street, Boston, MA 02115, USA
[35]International Hereditary Cancer Center, Department of Genetics and Pathology, Pomeranian Medical University, 70-115 Szczecin, Poland
[36]Division of Genetic Epidemiology, Department of Medicine, University of Utah School of Medicine, Salt Lake City, Utah 84112, USA
[37]George E. Wahlen Department of Veterans Affairs Medical Center, Salt Lake City, Utah 84148, USA
[38]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), D-69120, Heidelberg, Germany
[39]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), D-69120 Heidelberg, Germany
[40]Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Im Neuenheimer Feld 460
69120 Heidelberg, Germany
[41]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), D-69120 Heidelberg, Germany
[42]Network Aging Research, University of Heidelberg, Heidelberg, Germany
[43]Saarland Cancer Registry, D-66119 Saarbrücken, Germany
[44]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany
[45]Department of Cancer Epidemiology, Moffitt Cancer Center, 12902 Magnolia Drive, Tampa, FL 33612, USA
[46]School of Public Health, Louisiana State University Health Sciences Center, New Orleans, LA 70112, USA
[47]Department of Urology and Alexandrovska University Hospital, Medical University of Sofia, 1431 Sofia, Bulgaria
[48]Molecular Medicine Center, Department of Medical Chemistry and Biochemistry, Medical University of Sofia, Sofia, 2 Zdrave Str., 1431 Sofia, Bulgaria
[49]Institute of Health and Biomedical Innovation and School of Biomedical Sciences, Queensland University of Technology, Brisbane, QLD 4059, Australia
[50]Australian Prostate Cancer Research Centre-Qld, Translational Research Institute, Brisbane, Queensland 4102, Australia
[51]Australian Prostate Cancer Research Centre-Qld, Institute of Health and Biomedical Innovation and School of Biomedical Science, Queensland University of Technology, Brisbane QLD 4059, Australia
[52]Translational Research Institute, Brisbane, Queensland 4102, Australia
[53]Molecular Cancer Epidemiology Laboratory, QIMR Berghofer Institute of Medical Research, Brisbane, Australia
[54]Department of Genetics, Portuguese Oncology Institute of Porto (IPO-Porto),

4200-072 Porto, Portugal
[55]Biomedical Sciences Institute (ICBAS), University of Porto, 4050-313 Porto, Portugal
[56]Cancer Genetics Group, IPO-Porto Research Center (CI-IPOP), Portuguese Oncology Institute of Porto (IPO-Porto), Porto, Portugal
[57]The University of Surrey, Guildford, Surrey, GU2 7XH, UK
[58]Center for Cancer Research and Cell Biology, Queen's University of Belfast, Belfast, UK
[59]Nuffield Department of Surgical Sciences, John Radcliffe Hospital, University of Oxford, Oxford, UK
[60]NORMENT, KG Jebsen Centre, Oslo University Hospital and University of Oslo, Oslo, Norway
[61]Department of Radiology, University of California San Diego, La Jolla, CA, USA
[62]Department of Cognitive Science, University of California San Diego, La Jolla, CA, USA
[63]Department of Neurosciences, University of Californai San Diego, La Jolla, CA, USA

*Corresponding Author:

E-mail: rakarunamuni@ucsd.edu(RK), tseibert@ucsd.edu(TM)

^ Membership of The PRACTIAL Consortium is provided in the Supporting Information.

Abstract

We aimed to determine the effect of sample size on performance of polygenic hazard score (PHS) models in predicting the age at onset of prostate cancer. Age and genotypes were obtained for 40,861 men from the PRACTICAL consortium. The dataset included 201,590 SNPs per subject, and was split into training (34,444 samples) and testing (6,417 samples) sets. Two PHS model-building strategies were investigated. Established-SNP model considered 65 SNPs that had been associated with prostate cancer in the literature. A stepwise SNP selection was used to develop Discovery-SNP models. The performance of each PHS model was calculated for random sizes of the training set (1 to 30 thousand). The performance of a representative Established-SNP model was estimated for random sizes of the testing set (0.5 to 6 thousand). Mean $HR_{98/50}$ (hazard ratio of top 2% to the average in the test set) of the Established-SNP model increased from 1.73[95%CI: 1.69-1.77] to 2.41[2.40-2.43] when the number of training samples was increased from 1 to 30 thousand. The corresponding $HR_{98/50}$ of the Discovery-SNP model increased from 1.05[0.93-1.18] to 2.19[2.16-2.23]. $HR_{98/50}$ of a representative Established-SNP model using testing set sample sizes of 0.6 and 6 thousand observations were 1.78[1.70-1.85] and 1.73[1.71-1.76], respectively. We estimate that a study population of 20 to 30 thousand men is required to develop Discovery-SNP PHS models for prostate cancer. The required sample size could be reduced to 10 thousand samples, if a set of SNPs associated with the disease has already been established.

Author summary

Polygenic hazard scores represent a recent advancement in polygenic prediction to model the age of onset of various diseases, such as Alzheimer's disease or prostate cancer. These scores accumulate small effect sizes from several tens of genetic variants and can be used to establish an individual's risk of experiencing an event relative to a control population across time. The largest barrier to the development of polygenic hazard scores is the large number of study subjects needed to develop the underlying models. We sought to understand the effect of varying the total number of samples on the performance of a polygenic hazard score in the context of prostate cancer. We found that the performance of the score did not appreciably change beyond 20 to 30 thousand observations when developing the model from scratch. However, when the discovery of the genetic variants can be borrowed from those already identified in the literature to be associated with the disease, the required number of samples is reduced to 10 thousand with no appreciable detriment in performance. We hope that these results can guide the design of future studies of polygenic scores in other diseases and demonstrate the importance of genome-wide association studies.

Introduction

Polygenic prediction models have been studied extensively for several diseases such as prostate cancer[1], breast cancer[2], type 2 diabetes[3], dementia[4], and atherosclerosis[5]. Polygenic scores in the context of survival models are a more recent advancement in the field, but have been garnering interest in the prediction of age at onset of Alzheimer's disease[6] and prostate cancer[7]. The steady increase in genetic testing[8,9], both in public and clinical domains, suggests that survival models could be applied to new diseases. The largest obstacle to the development of these models is the large number of study subjects, often in the tens of thousands[8], which are required for robust training and testing.

Our aim was to quantify the effect of sample size on the performance of a polygenic survival model. This was explored through a specific disease condition that is expected to be representative, namely the prediction of age of onset in prostate cancer. We investigated two potential model development strategies. For the 'Established-SNP' model, we selected single-nucleotide polymorphisms (SNPs) that had previously been shown to be associated with prostate cancer, and simply estimated the coefficients for these SNPs in a Cox proportional hazards framework. For the 'Discovery-SNP' model, we implemented the SNP selection technique described by Seibert *et al.*[7] to identify SNPs in our genotyping data for inclusion in the Cox proportional hazards framework. The Established-SNP and Discovery-SNP represent two strategies that researchers could employ to build a polygenic survival model. In order to simulate samples of different sizes, we randomly sampled our training and testing sets. The results of this work will help inform the design of future studies to develop polygenic survival models for other diseases.

Results

Established- vs. Discovery-SNP model performance

Histogram comparisons of performance metrics of Established (EST) and Discovery (DIS) SNP models are illustrated in Figure 1. The performance metrics are shown for 50 random samplings of the training set using a sample size of 30 thousand total observations. Qualitatively, there appears to be more variability in performance metrics associated with the Discovery process.

Coefficients of Established-SNP model

The mean coefficients for the 65 SNPs used in the Established-SNP model are plotted in Figure 2.

Effect of training set sample size on performance

Box plots of the performance metrics of the Established-SNP and Discovery-SNP models for random samples of the training set are shown in Figure 3 and Figure 4, respectively. The mean values of $HR_{98/50}$, $HR_{20/50}$, $HR_{98/20}$, $HR_{80/20}$, z-score, and beta using a random training sample of 1 thousand total observations in the Established-SNP model were 1.73 [95% CI: 1.69-1.76], 0.71 [0.71-0.73], 2.42 [2.35-2.50], 1.96 [1.92-2.01], 9.92 [9.57-10.28], and 0.45 [0.43-

0.47] respectively. The corresponding values using a random training sample of 30 thousand total observations were 2.41 [95% CI: 2.40-2.43], 0.60 [0.60-0.60], 4.04 [4.02-4.07], 2.86 [2.84-2.87], 15.1 [15.04-15.16], and 1.18 [1.17-1.18] respectively.

The mean values of $HR_{98/50}$, $HR_{20/50}$, $HR_{98/20}$, $HR_{80/20}$, z-score, and beta using a random training sample of 1 thousand total observations in the Discovery-SNP model were 1.05 [0.93-1.18], 0.98 [0.89-1.07], 1.07 [0.91-1.24], 1.08 [0.91-1.24], 1.06 [-1.20-3.31], and 0.17 [-0.23-0.65] respectively. The corresponding performance values using a training sample size of 30 thousand observations were 2.20 [2.16-2.23], 1.60 [1.59-1.62], 3.47 [3.39-3.56], 2.53 [2.49-2.58], 13.19 [12.96-13.41], and 0.87 [0.85-0.89] respectively.

Effect of testing set sample size on performance

Box plots of the performance metrics of the representative Established-SNP model for random samples of the testing set are shown in Figure 5. The mean values of $HR_{98/50}$, $HR_{20/50}$, $HR_{98/20}$, $HR_{80/20}$, z-score, and beta using a random testing sample of 0.5 thousand total observations in the representative Established-SNP model were 1.78 [1.71-1.85], 0.73 [0.71-0.74], 2.50 [2.33-2.66], 1.99 [1.89-2.09], 3.82 [3.57-4.08], and 0.76 [0.70-0.82] respectively. The corresponding values using a testing sample of 6 thousand observations were: 1.73 [1.72-1.76], 0.73 [0.72-0.73], 2.39 [2.34-2.44], 1.93 [1.90-1.96], 13.07 [12.80-13.32], and 0.74 [0.72-0.76] respectively.

Discussion

We identified several trends in the effect of training and testing sample size on the performance of PHS models in predicting the age of onset of prostate cancer using SNP genetic variants. When using SNPs that had already been associated with prostate cancer risk, our analysis suggests that very little improvement in performance can be achieved once the training sets becomes larger than 10 to 15 thousand observations. When attempting to discover SNPs, a similar plateau in performance was observed from training sets larger than 20 to 25 thousand observations. Apart from z-scores, the performance metrics of the chosen Cox proportional hazards model did not vary with testing sample size. However, we did observe that the distribution of performance metrics narrows until a testing sample size of 3 to 4 thousand observations, after which the distribution remains relatively stable.

Our results may be used to inform researchers on the approximate number of subjects needed to develop PHS models to predict the age of onset of diseases using SNP counts. A dataset of 20 thousand observations may be the minimum needed to accurately estimate the PHS coefficients of SNPs that have been previously discovered in the setting of a logistic model. Such a dataset would allow for the accurate estimation of SNP coefficients as well as the testing of model performance in an independent holdout set. Based on our results, this number would have to be increased to roughly 30 thousand observations if the researchers intend on discovering the SNPs from scratch using the approach described here.

The PHS model developed by Desikan *et al.*[6] to predict age-associated risk of Alzheimer's disease used a training set with roughly 55,000 individuals. A similarly structured model developed by Seibert *et al.*[7] to guide screening for aggressive prostate cancer was developed with roughly 31,000 men. Studies such as these require large investments in time, money, and resources in order to acquire the genetic data needed for the analysis. The results of our analysis help elucidate that the minimum sample size needed to translate this technology to other diseases and processes may be lower than what has been used so far in previous studies. This seems to be particularly true if the researchers use SNPs that have already been discovered and validated as associated with the process of interest.

The results of this study must be considered in the context of its limitations. The list of Established-SNPs was previously selected from a larger dataset that included the sample patients used in the test set in the present study. As such, there is leakage of information from the test set to the development of the Established-SNP model. Therefore, the performance metrics of the Established-SNP model should not be directly compared to those of the Discovery-SNP model, as the values of the former may be inflated.

In addition, we have chosen to focus on only two of countless possible model development schemes. The role of sample size in other development strategies—such as regularized Cox proportional models, parametric survival functions, or random survival forests—is yet to be explored. Finally, the analysis is limited to prostate cancer and to the SNPs on the iCOGS array. Future work will include SNPs imputed from 1000 Genomes[13]. Such an analysis was not performed for this first study to limit computation time for bootstrap analyses and to avoid uncertainty due to imputation.

In conclusion, we have studied the effect of sample size on the performance of PHS models to study the association between SNPs and the age at onset of prostate cancer. We have determined that models require roughly 20 to 30 thousand samples before their performance would not be improved greatly by expansion of the training set. Using SNPs that have already been established in the literature may help reduce the number of training samples required to reach this performance plateau by almost 10 thousand samples.

Materials and Methods

Training and testing set

As previously described[7], we obtained genotype and age data from 21 studies included in the Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL) consortium. We analyzed data from 40,861 men consisting of 20,551 individuals with prostate cancer and 20,310 individuals without. For analysis, the age for each man was recorded as either their age at prostate cancer diagnosis (cases) or at interview (controls). Genotype data for 201,590 SNPs were also available for analysis. The genotype data had been assayed using a custom iCOGS chip (Illumina, San Diego, CA) the details for which are elaborated elsewhere[10]. The sample was split into training (34,444 men) and testing (6,417 men) sets. The testing set was selected using men who were enrolled in the Prostate testing for cancer and Treatment (ProtecT[11]) trial. ProtecT (ClinicalTrials.gov: NCT02044172) is a large, multicenter trial within the United Kingdom which aims to investigate the effectiveness of treatments for localized prostate cancer. The ProtecT study group was chosen for testing as it represented a well-characterized group of individuals that had been used for measuring testing performance for our earlier work. The Data Availability Statement describing how readers can gain access to the PRACTICAL dataset is provided in the Supplementary Information.

Established-SNP model

A list of 65 SNPs[12] was chosen to represent those on the iCOGS array that had been published as associated with prostate cancer. The coefficients of the SNPs within the Established-SNP model were then estimated using the "coxphfit" function in MATLAB (Mathworks, Natwick, MA). Prior to parameter estimation, missing SNP data were replaced by mean imputation. It should be noted that the 65 SNPs used were discovered, in large part, using the data presently defined as the test set. The effect allele for all 65 SNPs was defined as "A" to simplify analysis.

Discovery-SNP model

SNPs with call rates less than 95% were removed from the selection process. For every SNP, a trend test was used to check for associations between SNP count and the binary classification of individuals with or without prostate cancer. The SNP selection pool was then reduced to those whose trend test p-value was less 1x10^{-6}. In order of increasing p-value, each SNP was tested in a multiple logistic regression model for association with the binary classification of men as with or without prostate cancer, after adjusting for age, six principal components based upon genetic ancestry, and previously selected SNPs. If the p-value of the coefficient of the tested SNP was less than 1x10^{-6}, it was selected for the final Cox proportional hazard model estimation. The coefficients of the selected SNP pool within the Discovery-SNP model were estimated as previously described[7].

Polygenic Hazard Score (PHS)

The polygenic hazard score (PHS) for each of the Established-SNP and Discovery-SNP models was calculated as the linear product of the coefficients of the SNPs used in the model and the corresponding patient genotype counts[6,7].

PHS performance metrics

Several performance metrics for PHS models were investigated, and are described in Table 1. In each case, the PHS for each test subject was calculated as the dot product of SNP coefficients, either Established or Discovery, and SNP counts. A Cox proportional hazards model was then fit using PHS as the sole predictor of age in the test set. The z-score and beta of this Cox proportional hazards model relate to how well PHS was associated with age within the test set. The hazard ratios were calculated as the exponential of the differences in predicted log-relative hazards of different groups within the test set. The groups were defined using centile cut-points for those controls within the training set whose age was less than 70 years. This list of performance metrics expands on those (z-score and $HR_{98/50}$) that were used in our earlier work[7].

Table 1. Performance metrics used in the evaluation of polygenic hazard scores.

Performance metric	Description
$HR_{98/50}$	Hazard ratio of the top 2% to the average (30 – 70%) in the test set
$HR_{20/50}$	Hazard ratio of the bottom 20% to the average (30 – 70%) in the test set
$HR_{98/20}$	Hazard ratio of the top 2% to the bottom 20% in the test set
$HR_{80/20}$	Hazard ratio of the top 20% to the bottom 20% in the test set.
z-score	z-score of Cox proportional hazards model using PHS as a sole predictor of age in the test set
beta	coefficient of PHS in a Cox proportional hazards model using PHS as a sole predictor of age in the test set.

Random sampling of training set

Random sampling of the training set was performed with replacement while ensuring equal proportions of men with and without prostate cancer. The training set was randomly sampled to include 1, 5, 10, 15, 20, 25, and 30 thousand total observations. Performance of the Established and Discovery-SNP models using random samples of the training data was measured in the entire test set.

Random sampling of the testing set

Random sampling of the testing set was performed with replacement while ensuring equal proportion of men with and without prostate cancer. The testing set was randomly sampled to include 0.5, 1, 2, 3, 4, 5 and 6 thousand total observations. Performance in the randomly sampled testing sets was performed using a representative Established-SNP model. The representative model was chosen as that whose parameters were estimated using a training sample size of 30 thousand total observations, and whose performance metrics were the shortest Euclidean distance to the average performance across all Established-SNP models using a training sample size of 30 thousand.

References

1. 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;60: 21–28. doi:10.1016/j.eururo.2011.01.017
2. Machiela MJ, Chen C, Chanock SJ, Hunter DJ, Kraft P. Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. Genet Epidemiol. 2011;514: n/a-n/a. doi:10.1002/gepi.20600
3. Vassy JL, Hivert MF, Porneala B, Dauriz M, Florez JC, Dupuis J, et al. Polygenic type 2 diabetes prediction at the limit of common variant detection. Diabetes. 2014;63: 2172–2182. doi:10.2337/db13-1663
4. Marden JR, Walter S, Tchetgen Tchetgen EJ, Kawachi I, Glymour MM. Validation of a polygenic risk score for dementia in black and white individuals. Brain Behav. 2014;4: 687–697. doi:10.1002/brb3.248
5. Natarajan P, Young R, Stitziel NO, Padmanabhan S, Baber U, Mehran R, et al. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. Circulation. 2017;135: 2091–2101. doi:10.1161/CIRCULATIONAHA.116.024436
6. Desikan RS, Fan CC, Wang Y, Schork AJ, Cabral HJ, Cupples LA, et al. Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. PLoS Med. 2017;14: 1–17. doi:10.1371/journal.pmed.1002258

7. 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. doi:10.1136/bmj.j5757

8. Chatterjee N, Shi J, García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat Rev Genet. Nature Publishing Group; 2016;17: 392–406. doi:10.1038/nrg.2016.27

9. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. Springer US; 2018;19: 581–590. doi:10.1038/s41576-018-0018-x

10. 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. doi:10.1038/ng.2560

11. Lane JA, Donovan JL, Davis M, Walsh E, Dedman D, Down L, et al. Active monitoring, radical prostatectomy, or radiotherapy for localised prostate cancer: Study design and diagnostic and baseline results of the ProtecT randomised phase 3 trial. Lancet Oncol. Lane et al. Open Access article distributed under the terms of CC BY; 2014;15: 1109–1118. doi:10.1016/S1470-2045(14)70361-4

12. Szulkin R, Whitington T, Eklund M, Aly M, Eeles RA, Easton D, et al. Prediction of individual genetic risk to prostate cancer using a polygenic score. Prostate. 2015;75: 1467–1474. doi:10.1002/pros.23037

13. Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467: 1061–1073. doi:10.1038/nature09534

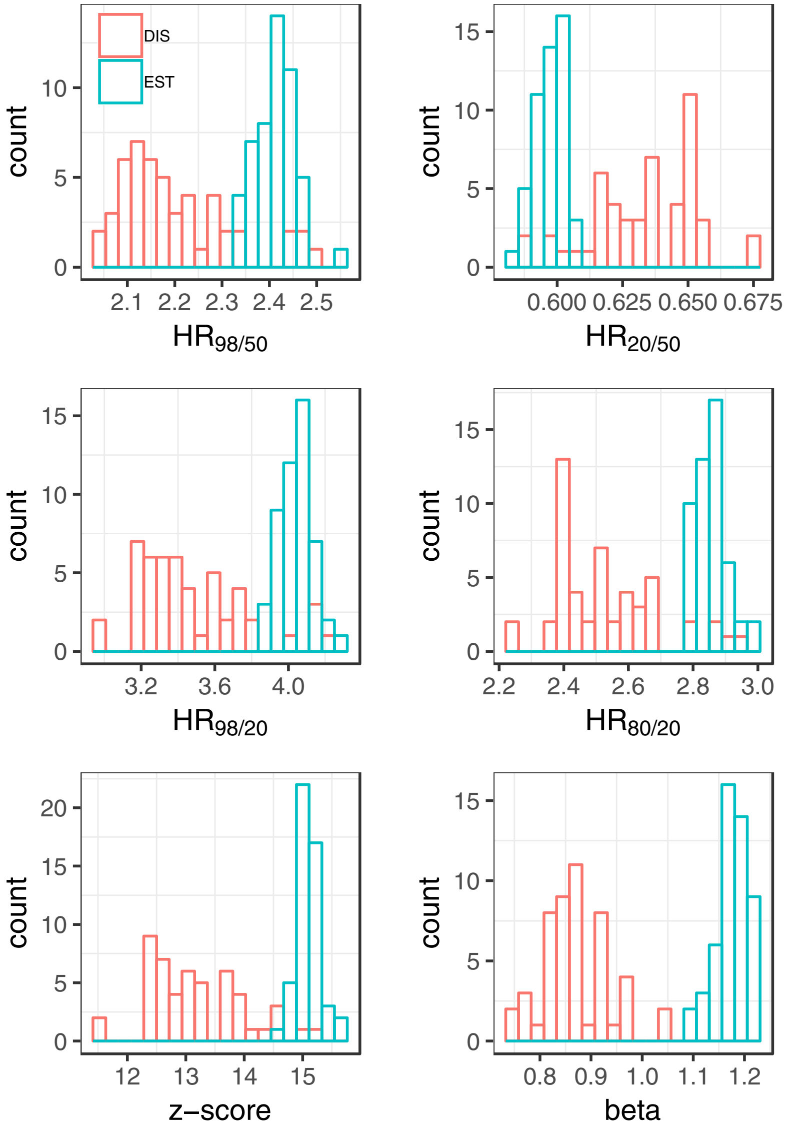

Figure 1. Comparison of performance metrics between Established (EST) and Discovery (DIS) SNP models using 50 random samples of the training set using a sample size of 30 thousand. There is more variability with the Discovery process. Established SNPs, though, were discovered using the data in the training set; this circularity is not accounted for in the present study, which focuses on sample size effects.

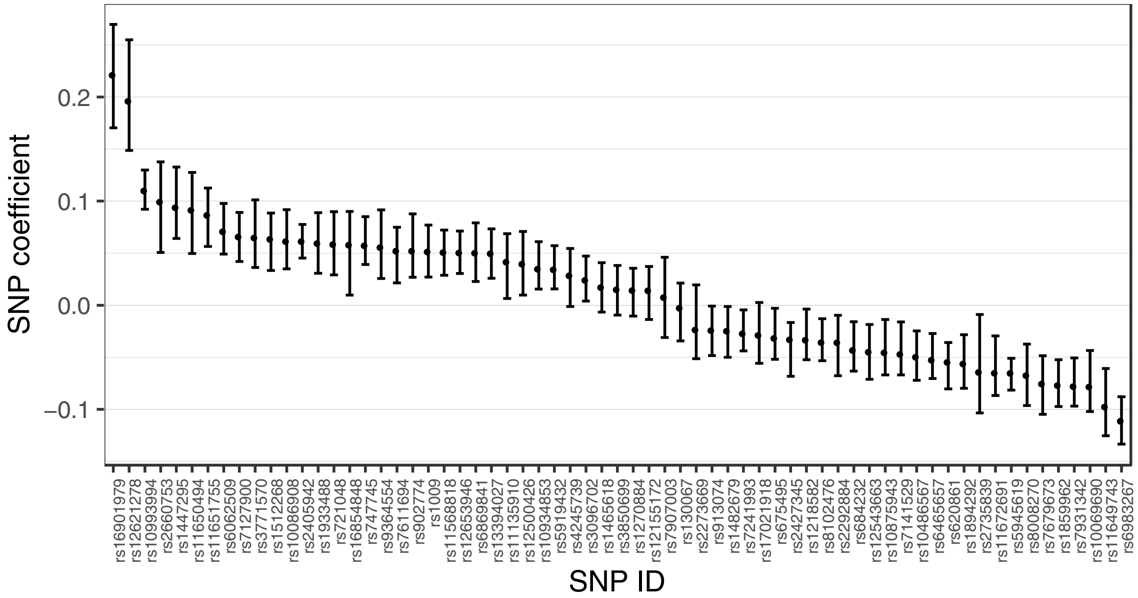

Figure 2. Coefficients of 65 SNPs used in the Established SNP model. Data points represent mean values across 50 iterations of a random sample of the training set using a sample size of 30 thousand total observations. Error bars represent 95% confidence intervals.

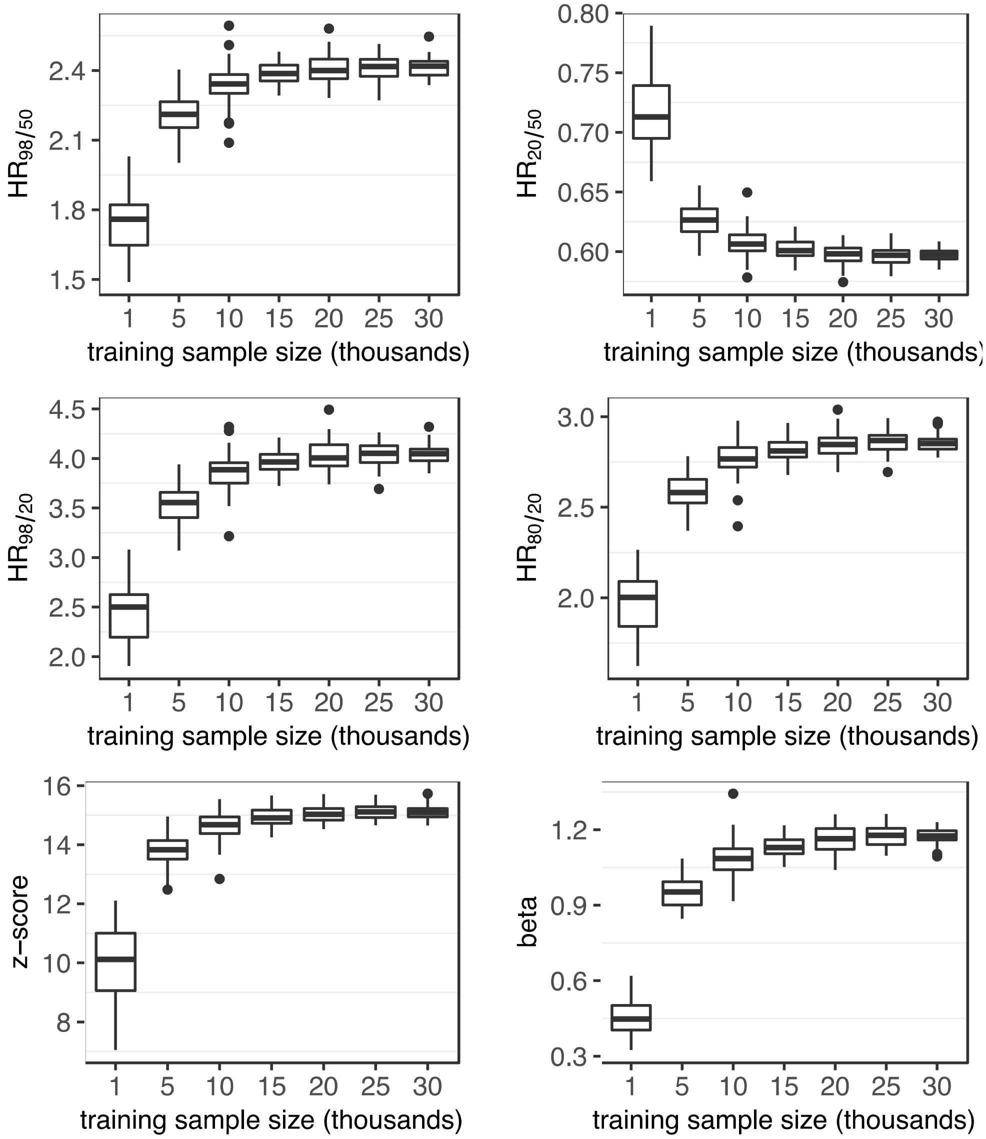

Figure 3. Performance metrics of Established SNP model. Box plots of performance metrics are shown for random samples of the training set using sample sizes of 1, 5, 10, 15, 20, 25, and 30 thousand total observations. Within each box plot, the horizontal line represents the median and the box extends from the 25th to 75th percentile.

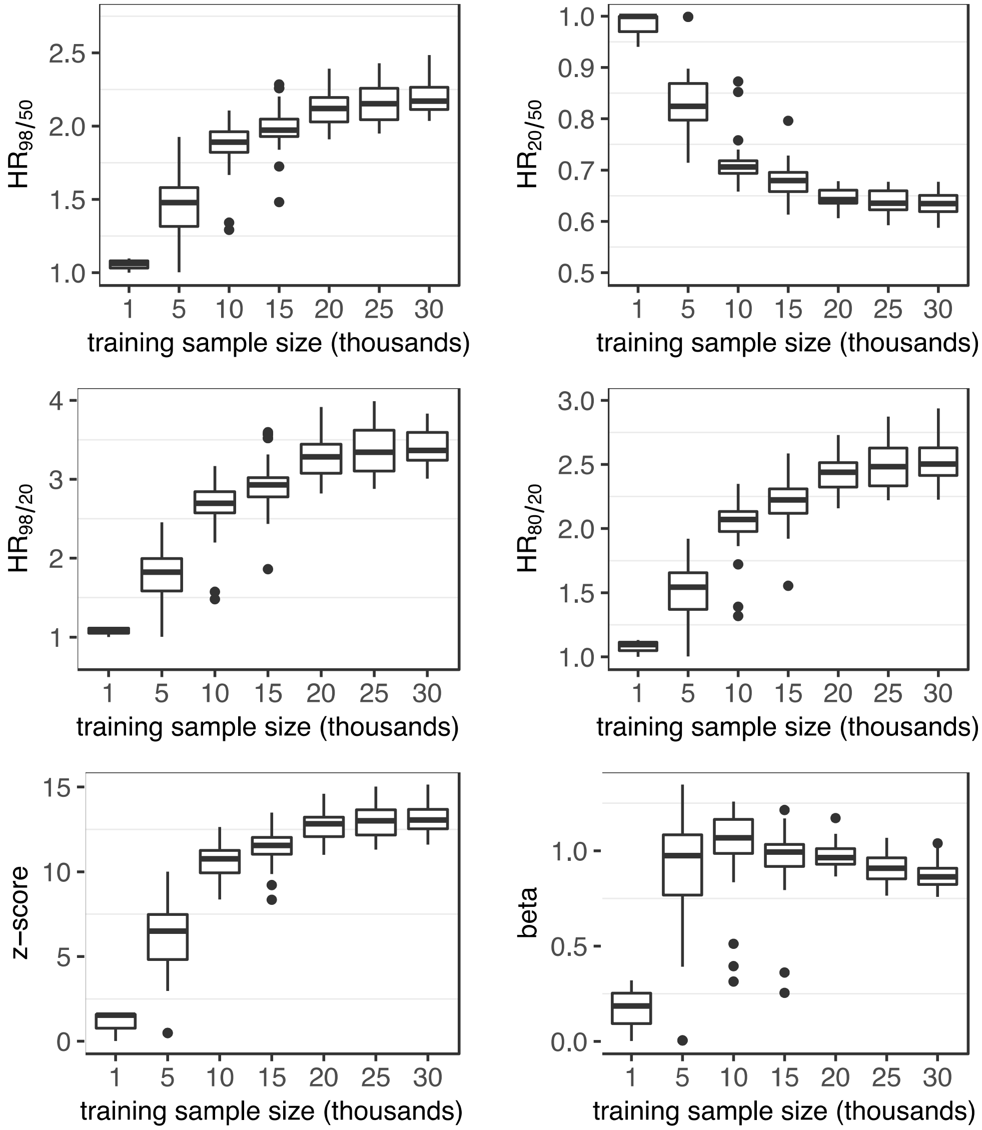

Figure 4. Performance metrics of the Discovery SNP model. Box plots of performance metrics are shown for random samples of the training set using sample sizes of 1, 5, 10, 15, 20, 25, and 30 thousand total observations. Within each box plot, the horizontal line represents the median and the box extends from the 25th to 75th percentile.

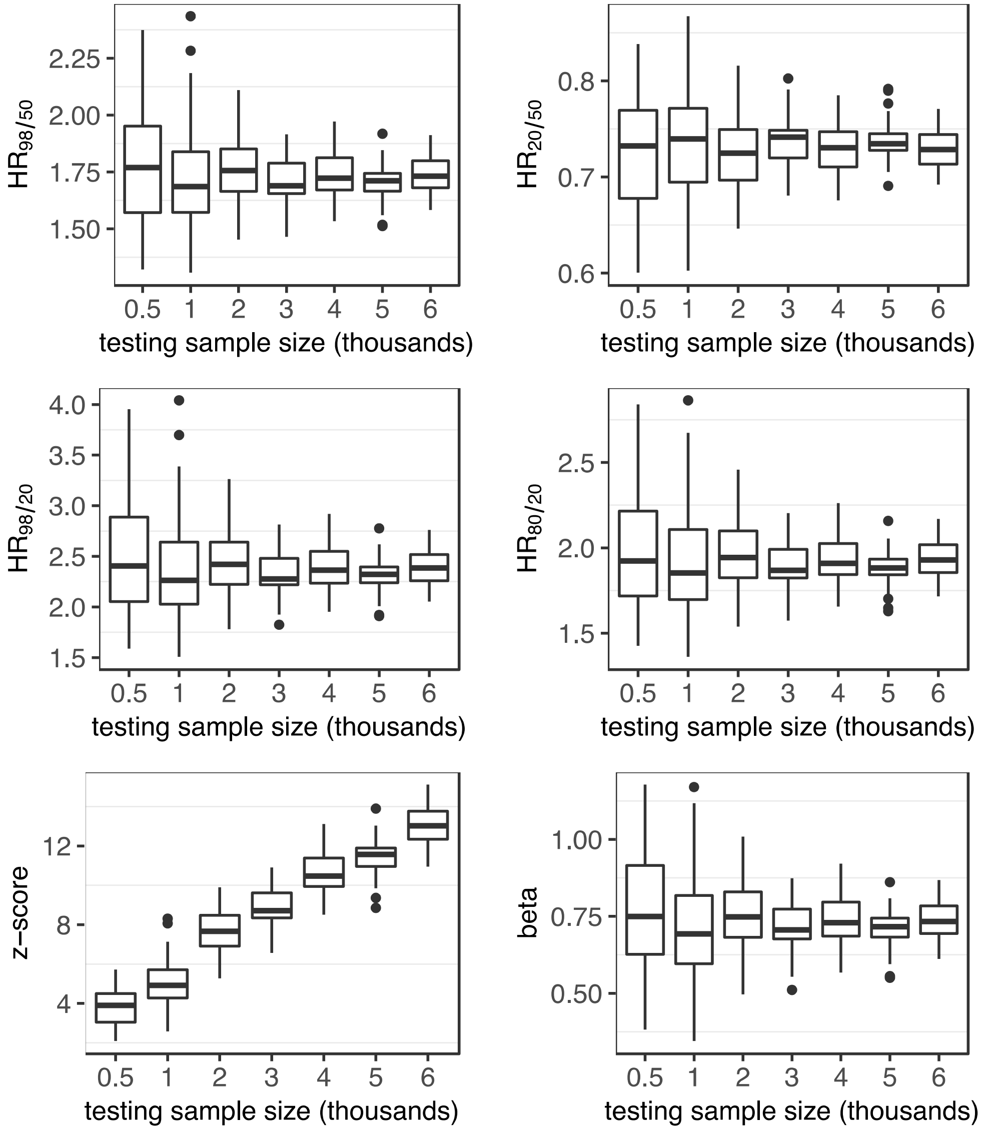

Figure 5. Performance as a function of testing sample size. Box plots of performance metrics of the representative Established SNP model in random samples of the testing set from 0.5 to 6 thousand total observations.

Supporting Information Legends

Supporting Information 1. Data Availability Statement details how readers can obtain the data from the PRACTICAL (Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome) consortium. The document also contains the additional authorship, affiliation, and funding sources for the PRACTICAL consortium.