

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

2

3 Roshan A. Karunamuni¹, Minh-Phuong Huynh-Le¹, Chun C. Fan², Rosalind A.

4 Eeles^{3,4}, Douglas F. Easton⁵, ZSofia Kote-Jarai³, Ali Amin Al Olama^{5,6}, Sara

5 Benlloch Garcia⁵, Kenneth Muir^{7,8}, Henrik Gronberg⁹, Fredrik Wiklund⁹, Markus

6 Aly^{9,10,11}, Johanna Schleutker^{12,13}, Csilla Sipeky¹², Teuvo LJ Tammela^{14,15}, Børge

7 G. Nordestgaard^{16,17}, Tim J. Key¹⁸, Ruth C. Travis¹⁸, David E. Neal^{19,20,21}, Jenny

8 L. Donovan²², Freddie C. Hamdy^{23,24}, Paul Pharoah²⁵, Nora Pashayan^{26,25,27},

9 Kay-Tee Khaw²⁸, Stephen N. Thibodeau²⁹, Shannon K. McDonnell³⁰, Daniel J.

10 Schaid³⁰, Christiane Maier³¹, Walther Vogel³², Manuel Luedeke³¹, Kathleen

11 Herkommer³³, Adam S. Kibel³⁴, Cezary Cybulski³⁵, Dominika Wokolorczyk³⁵,

12 Wojciech Kluzniak³⁵, Lisa Cannon-Albright^{36,37}, Hermann Brenner^{38,39,40}, Ben

13 Schöttker^{41,42}, Bernd Holleczeck^{43,44}, Jong Y. Park⁴⁵, Thomas A. Sellers⁴⁵, Hui-Yi

14 Lin⁴⁶, Chavdar Slavov⁴⁷, Radka Kaneva⁴⁸, Vanio Mitev⁴⁸, Jyotsna Batra^{49,50},

15 Judith A. Clements^{51,52}, Amanda Spurdle⁵³, Australian Prostate Cancer

16 BioResource (APCB)⁵¹, Manuel R. Teixeira^{54,55}, Paula Paulo^{54,56}, Sofia Maia^{54,56},

17 Hardev Pandha⁵⁷, Agnieszka Michael⁵⁷, Ian G. Mills^{58,59}, Ole A. Andreassen⁶⁰,

18 Anders M. Dale^{61,62,63}, Tyler M. Seibert¹, The PRACTICAL Consortium[^]

19 ¹Department of Radiation Medicine and Applied Sciences, University of California
20 San Diego, La Jolla, CA, USA

21 ²Healthlytix, 4747 Executive Dr. Suite 820, San Diego, CA, USA

22 ³The Institute of Cancer Research, London, SM2 5NG, UK

23 ⁴Royal Marsden NHS Foundation Trust, London, SW3 6JJ, UK

24 ⁵Centre for Cancer Genetic Epidemiology, Department of Public Health and
25 Primary Care, University of Cambridge, Strangeways Research Laboratory,
26 Cambridge CB1 8RN, UK

27 ⁶University of Cambridge, Department of Clinical Neurosciences, Stroke
28 Research Group, R3, Box 83, Cambridge Biomedical Campus, Cambridge CB2

29 0QQ, UK

30 ⁷Division of Population Health, Health Services Research and Primary Care,
31 University of Manchester, Oxford Road, Manchester, M13 9PL, UK

32 ⁸Warwick Medical School, University of Warwick, Coventry, UK

33 ⁹Department of Medical Epidemiology and Biostatistics, Karolinska Institute, SE-
34 171 77 Stockholm, Sweden

35 ¹⁰Department of Molecular Medicine and Surgery, Karolinska Institute, SE-171
36 77 Stockholm, Sweden

37 ¹¹Department of Urology, Karolinska University Hospital, Stockholm, Sweden

38 ¹²Institute of Biomedicine, Kiinamyllynkatu 10, FI-20014 University of Turku,
39 Finland

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

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

44 ¹⁵Department of Urology, Tampere University Hospital, Tampere, Finland

45 ¹⁶Faculty of Health and Medical Sciences, University of Copenhagen, 2200
46 Copenhagen, Denmark

47 ¹⁷Department of Clinical Biochemistry, Herlev and Gentofte Hospital,
48 Copenhagen University Hospital, Herlev, 2200 Copenhagen, Denmark

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

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

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

56 ²¹Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre,
57 Cambridge UK

58 ²²School of Social and Community Medicine, University of Bristol, Canynge Hall,
59 39 Whatley Road, Bristol, BS8 2PS, UK

60 ²³Nuffield Department of Surgical Sciences, University of Oxford, Oxford, OX1
61 2JD, UK

62 ²⁴Faculty of Medical Science, University of Oxford, John Radcliffe Hospital,
63 Oxford, UK

64 ²⁵Centre for Cancer Genetic Epidemiology, Department of Oncology, University
65 of Cambridge, Strangeways Laboratory, Worts Causeway, Cambridge, CB1
66 8RN, UK

67 ²⁶University College London, Department of Applied Health Research,
68 London, UK

69 ²⁷Department of Applied Health Research, University College London, London,
70 WC1E 7HB, UK

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

72 ²⁹Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester,
73 MN 55905, USA

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

- 75 55905, USA
- 76 ³¹Humangenetik Tuebingen, Paul-Ehrlich-Str 23, D-72076 Tuebingen
- 77 ³²Institute for Human Genetics, University Hospital Ulm, 89075 Ulm, Germany
- 78 ³³Technical University of Munich, School of Medicine, Klinikum rechts der Isar,
- 79 Department of Urology
- 80 ³⁴Division of Urologic Surgery, Brigham and Womens Hospital, 75 Francis Street,
- 81 Boston, MA 02115, USA
- 82 ³⁵International Hereditary Cancer Center, Department of Genetics and Pathology,
- 83 Pomeranian Medical University, 70-115 Szczecin, Poland
- 84 ³⁶Division of Genetic Epidemiology, Department of Medicine, University of Utah
- 85 School of Medicine, Salt Lake City, Utah 84112, USA
- 86 ³⁷George E. Wahlen Department of Veterans Affairs Medical Center, Salt Lake
- 87 City, Utah 84148, USA
- 88 ³⁸Division of Clinical Epidemiology and Aging Research, German Cancer
- 89 Research Center (DKFZ), D-69120, Heidelberg, Germany
- 90 ³⁹German Cancer Consortium (DKTK), German Cancer Research Center
- 91 (DKFZ), D-69120 Heidelberg, Germany
- 92 ⁴⁰Division of Preventive Oncology, German Cancer Research Center (DKFZ) and
- 93 National Center for Tumor Diseases (NCT), Im Neuenheimer Feld 460
- 94 69120 Heidelberg, Germany
- 95 ⁴¹Division of Clinical Epidemiology and Aging Research, German Cancer
- 96 Research Center (DKFZ), D-69120 Heidelberg, Germany
- 97 ⁴²Network Aging Research, University of Heidelberg, Heidelberg, Germany
- 98 ⁴³Saarland Cancer Registry, D-66119 Saarbrücken, Germany
- 99 ⁴⁴Division of Clinical Epidemiology and Aging Research, German Cancer
- 100 Research Center (DKFZ), Heidelberg, Germany
- 101 ⁴⁵Department of Cancer Epidemiology, Moffitt Cancer Center, 12902 Magnolia
- 102 Drive, Tampa, FL 33612, USA
- 103 ⁴⁶School of Public Health, Louisiana State University Health Sciences Center,
- 104 New Orleans, LA 70112, USA
- 105 ⁴⁷Department of Urology and Alexandrovska University Hospital, Medical
- 106 University of Sofia, 1431 Sofia, Bulgaria
- 107 ⁴⁸Molecular Medicine Center, Department of Medical Chemistry and
- 108 Biochemistry, Medical University of Sofia, Sofia, 2 Zdrave Str., 1431 Sofia,
- 109 Bulgaria
- 110 ⁴⁹Institute of Health and Biomedical Innovation and School of Biomedical
- 111 Sciences, Queensland University of Technology, Brisbane, QLD 4059, Australia
- 112 ⁵⁰Australian Prostate Cancer Research Centre-Qld, Translational Research
- 113 Institute, Brisbane, Queensland 4102, Australia
- 114 ⁵¹Australian Prostate Cancer Research Centre-Qld, Institute of Health and
- 115 Biomedical Innovation and School of Biomedical Science, Queensland University
- 116 of Technology, Brisbane QLD 4059, Australia
- 117 ⁵²Translational Research Institute, Brisbane, Queensland 4102, Australia
- 118 ⁵³Molecular Cancer Epidemiology Laboratory, QIMR Berghofer Institute of
- 119 Medical Research, Brisbane, Australia
- 120 ⁵⁴Department of Genetics, Portuguese Oncology Institute of Porto (IPO-Porto),

121 4200-072 Porto, Portugal

122 ⁵⁵Biomedical Sciences Institute (ICBAS), University of Porto, 4050-313 Porto,
123 Portugal

124 ⁵⁶Cancer Genetics Group, IPO-Porto Research Center (CI-IPOP), Portuguese
125 Oncology Institute of Porto (IPO-Porto), Porto, Portugal

126 ⁵⁷The University of Surrey, Guildford, Surrey, GU2 7XH, UK

127 ⁵⁸Center for Cancer Research and Cell Biology, Queen's University of Belfast,
128 Belfast, UK

129 ⁵⁹Nuffield Department of Surgical Sciences, John Radcliffe Hospital, University of
130 Oxford, Oxford, UK

131 ⁶⁰NORMENT, KG Jebsen Centre, Oslo University Hospital and University of
132 Oslo, Oslo, Norway

133 ⁶¹Department of Radiology, University of California San Diego, La Jolla, CA, USA

134 ⁶²Department of Cognitive Science, University of California San Diego, La Jolla,
135 CA, USA

136 ⁶³Department of Neurosciences, University of California San Diego, La Jolla, CA,
137 USA

138

139

140 *Corresponding Author:

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

142

143 ^ Membership of The PRACTIAL Consortium is provided in the Supporting

144 Information.

145

146 **Abstract**

147 We aimed to determine the effect of sample size on performance of polygenic
148 hazard score (PHS) models in predicting the age at onset of prostate cancer.
149 Age and genotypes were obtained for 40,861 men from the PRACTICAL
150 consortium. The dataset included 201,590 SNPs per subject, and was split into
151 training (34,444 samples) and testing (6,417 samples) sets. Two PHS model-
152 building strategies were investigated. Established-SNP model considered 65
153 SNPs that had been associated with prostate cancer in the literature. A stepwise
154 SNP selection was used to develop Discovery-SNP models. The performance of
155 each PHS model was calculated for random sizes of the training set (1 to 30
156 thousand). The performance of a representative Established-SNP model was
157 estimated for random sizes of the testing set (0.5 to 6 thousand). Mean $HR_{98/50}$
158 (hazard ratio of top 2% to the average in the test set) of the Established-SNP
159 model increased from 1.73[95%CI: 1.69-1.77] to 2.41[2.40-2.43] when the
160 number of training samples was increased from 1 to 30 thousand. The
161 corresponding $HR_{98/50}$ of the Discovery-SNP model increased from 1.05[0.93-
162 1.18] to 2.19[2.16-2.23]. $HR_{98/50}$ of a representative Established-SNP model using
163 testing set sample sizes of 0.6 and 6 thousand observations were 1.78[1.70-1.85]
164 and 1.73[1.71-1.76], respectively. We estimate that a study population of 20 to 30
165 thousand men is required to develop Discovery-SNP PHS models for prostate
166 cancer. The required sample size could be reduced to 10 thousand samples, if a
167 set of SNPs associated with the disease has already been established.

168

169 **Author summary**

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

186 **Introduction**

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

197 Our aim was to quantify the effect of sample size on the performance of a
198 polygenic survival model. This was explored through a specific disease condition
199 that is expected to be representative, namely the prediction of age of onset in
200 prostate cancer. We investigated two potential model development strategies.
201 For the 'Established-SNP' model, we selected single-nucleotide polymorphisms
202 (SNPs) that had previously been shown to be associated with prostate cancer,
203 and simply estimated the coefficients for these SNPs in a Cox proportional
204 hazards framework. For the 'Discovery-SNP' model, we implemented the SNP
205 selection technique described by Seibert *et al.*[7] to identify SNPs in our
206 genotyping data for inclusion in the Cox proportional hazards framework. The
207 Established-SNP and Discovery-SNP represent two strategies that researchers
208 could employ to build a polygenic survival model. In order to simulate samples of

209 different sizes, we randomly sampled our training and testing sets. The results of
210 this work will help inform the design of future studies to develop polygenic
211 survival models for other diseases.

212

213 **Results**

214 Established- vs. Discovery-SNP model performance

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

220

221 Coefficients of Established-SNP model

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

224

225 Effect of training set sample size on performance

226 Box plots of the performance metrics of the Established-SNP and
227 Discovery-SNP models for random samples of the training set are shown in
228 Figure 3 and Figure 4, respectively. The mean values of $HR_{98/50}$, $HR_{20/50}$, $HR_{98/20}$,
229 $HR_{80/20}$, z-score, and beta using a random training sample of 1 thousand total
230 observations in the Established-SNP model were 1.73 [95% CI: 1.69-1.76], 0.71
231 [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-

232 0.47] respectively. The corresponding values using a random training sample of
233 30 thousand total observations were 2.41 [95% CI: 2.40-2.43], 0.60 [0.60-0.60],
234 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]
235 respectively.

236 The mean values of $HR_{98/50}$, $HR_{20/50}$, $HR_{98/20}$, $HR_{80/20}$, z-score, and beta
237 using a random training sample of 1 thousand total observations in the
238 Discovery-SNP model were 1.05 [0.93-1.18], 0.98 [0.89-1.07], 1.07 [0.91-1.24],
239 1.08 [0.91-1.24], 1.06 [-1.20-3.31], and 0.17 [-0.23-0.65] respectively. The
240 corresponding performance values using a training sample size of 30 thousand
241 observations were 2.20 [2.16-2.23], 1.60 [1.59-1.62], 3.47 [3.39-3.56], 2.53 [2.49-
242 2.58], 13.19 [12.96-13.41], and 0.87 [0.85-0.89] respectively.

243

244 Effect of testing set sample size on performance

245 Box plots of the performance metrics of the representative Established-
246 SNP model for random samples of the testing set are shown in Figure 5. The
247 mean values of $HR_{98/50}$, $HR_{20/50}$, $HR_{98/20}$, $HR_{80/20}$, z-score, and beta using a
248 random testing sample of 0.5 thousand total observations in the representative
249 Established-SNP model were 1.78 [1.71-1.85], 0.73 [0.71-0.74], 2.50 [2.33-2.66],
250 1.99 [1.89-2.09], 3.82 [3.57-4.08], and 0.76 [0.70-0.82] respectively. The
251 corresponding values using a testing sample of 6 thousand observations were:
252 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
253 [12.80-13.32], and 0.74 [0.72-0.76] respectively.

254

255 **Discussion**

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

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

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

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

296 In addition, we have chosen to focus on only two of countless possible
297 model development schemes. The role of sample size in other development
298 strategies—such as regularized Cox proportional models, parametric survival
299 functions, or random survival forests—is yet to be explored. Finally, the analysis
300 is limited to prostate cancer and to the SNPs on the iCOGS array. Future work

301 will include SNPs imputed from 1000 Genomes[13]. Such an analysis was not
302 performed for this first study to limit computation time for bootstrap analyses and
303 to avoid uncertainty due to imputation.

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

311

312 **Materials and Methods**

313 Training and testing set

314 As previously described[7], we obtained genotype and age data from 21
315 studies included in the Prostate Cancer Association Group to Investigate Cancer
316 Associated Alterations in the Genome (PRACTICAL) consortium. We analyzed
317 data from 40,861 men consisting of 20,551 individuals with prostate cancer and
318 20,310 individuals without. For analysis, the age for each man was recorded as
319 either their age at prostate cancer diagnosis (cases) or at interview (controls).
320 Genotype data for 201,590 SNPs were also available for analysis. The genotype
321 data had been assayed using a custom iCOGS chip (Illumina, San Diego, CA)
322 the details for which are elaborated elsewhere[10]. The sample was split into
323 training (34,444 men) and testing (6,417 men) sets. The testing set was selected

324 using men who were enrolled in the Prostate testing for cancer and Treatment
325 (ProtecT[11]) trial. ProtecT (ClinicalTrials.gov: NCT02044172) is a large,
326 multicenter trial within the United Kingdom which aims to investigate the
327 effectiveness of treatments for localized prostate cancer. The ProtecT study
328 group was chosen for testing as it represented a well-characterized group of
329 individuals that had been used for measuring testing performance for our earlier
330 work. The Data Availability Statement describing how readers can gain access to
331 the PRACTICAL dataset is provided in the Supplementary Information.

332

333 Established-SNP model

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

342

343 Discovery-SNP model

344 SNPs with call rates less than 95% were removed from the selection
345 process. For every SNP, a trend test was used to check for associations between
346 SNP count and the binary classification of individuals with or without prostate

347 cancer. The SNP selection pool was then reduced to those whose trend test p-
348 value was less 1×10^{-6} . In order of increasing p-value, each SNP was tested in a
349 multiple logistic regression model for association with the binary classification of
350 men as with or without prostate cancer, after adjusting for age, six principal
351 components based upon genetic ancestry, and previously selected SNPs. If the
352 p-value of the coefficient of the tested SNP was less than 1×10^{-6} , it was selected
353 for the final Cox proportional hazard model estimation. The coefficients of the
354 selected SNP pool within the Discovery-SNP model were estimated as previously
355 described[7].

356

357 Polygenic Hazard Score (PHS)

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

361

362 PHS performance metrics

363 Several performance metrics for PHS models were investigated, and are
364 described in Table 1. In each case, the PHS for each test subject was calculated
365 as the dot product of SNP coefficients, either Established or Discovery, and SNP
366 counts. A Cox proportional hazards model was then fit using PHS as the sole
367 predictor of age in the test set. The z-score and beta of this Cox proportional
368 hazards model relate to how well PHS was associated with age within the test
369 set. The hazard ratios were calculated as the exponential of the differences in

370 predicted log-relative hazards of different groups within the test set. The groups
371 were defined using centile cut-points for those controls within the training set
372 whose age was less than 70 years. This list of performance metrics expands on
373 those (z-score and $HR_{98/50}$) that were used in our earlier work[7].

374

375 **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.

376

377 Random sampling of training set

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

384

385 Random sampling of the testing set

386 Random sampling of the testing set was performed with replacement while
387 ensuring equal proportion of men with and without prostate cancer. The testing
388 set was randomly sampled to include 0.5, 1, 2, 3, 4, 5 and 6 thousand total
389 observations. Performance in the randomly sampled testing sets was performed
390 using a representative Established-SNP model. The representative model was
391 chosen as that whose parameters were estimated using a training sample size of
392 30 thousand total observations, and whose performance metrics were the
393 shortest Euclidean distance to the average performance across all Established-
394 SNP models using a training sample size of 30 thousand.

395

396

397 **References**

- 398 1. Aly M, Wiklund F, Xu J, Isaacs WB, Eklund M, D'Amato M, et al. Polygenic
399 risk score improves prostate cancer risk prediction: Results from the
400 Stockholm-1 cohort study. *Eur Urol.* 2011;60: 21–28.
401 doi:10.1016/j.eururo.2011.01.017
- 402 2. Machiela MJ, Chen C, Chanock SJ, Hunter DJ, Kraft P. Evaluation of
403 polygenic risk scores for predicting breast and prostate cancer risk. *Genet
404 Epidemiol.* 2011;514: n/a-n/a. doi:10.1002/gepi.20600
- 405 3. Vassy JL, Hivert MF, Porneala B, Dauriz M, Florez JC, Dupuis J, et al.
406 Polygenic type 2 diabetes prediction at the limit of common variant
407 detection. *Diabetes.* 2014;63: 2172–2182. doi:10.2337/db13-1663
- 408 4. Marden JR, Walter S, Tchetgen Tchetgen EJ, Kawachi I, Glymour MM.
409 Validation of a polygenic risk score for dementia in black and white
410 individuals. *Brain Behav.* 2014;4: 687–697. doi:10.1002/brb3.248
- 411 5. Natarajan P, Young R, Stitzel NO, Padmanabhan S, Baber U, Mehran R,
412 et al. Polygenic risk score identifies subgroup with higher burden of
413 atherosclerosis and greater relative benefit from statin therapy in the
414 primary prevention setting. *Circulation.* 2017;135: 2091–2101.
415 doi:10.1161/CIRCULATIONAHA.116.024436
- 416 6. Desikan RS, Fan CC, Wang Y, Schork AJ, Cabral HJ, Cupples LA, et al.
417 Genetic assessment of age-associated Alzheimer disease risk:
418 Development and validation of a polygenic hazard score. *PLoS Med.*
419 2017;14: 1–17. doi:10.1371/journal.pmed.1002258

- 420 7. Seibert TM, Fan CC, Wang Y, Zuber V, Karunamuni R, Parsons JK, et al.
421 Polygenic hazard score to guide screening for aggressive prostate cancer:
422 Development and validation in large scale cohorts. *BMJ*. 2018;360: 1–7.
423 doi:10.1136/bmj.j5757
- 424 8. Chatterjee N, Shi J, García-Closas M. Developing and evaluating polygenic
425 risk prediction models for stratified disease prevention. *Nat Rev Genet*.
426 Nature Publishing Group; 2016;17: 392–406. doi:10.1038/nrg.2016.27
- 427 9. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of
428 polygenic risk scores. *Nat Rev Genet*. Springer US; 2018;19: 581–590.
429 doi:10.1038/s41576-018-0018-x
- 430 10. Eeles RA, Olama AA Al, Benlloch S, Saunders EJ, Leongamornlert DA,
431 Tymrakiewicz M, et al. Identification of 23 new prostate cancer
432 susceptibility loci using the iCOGS custom genotyping array. *Nat Genet*.
433 2013;45: 385–391. doi:10.1038/ng.2560
- 434 11. Lane JA, Donovan JL, Davis M, Walsh E, Dedman D, Down L, et al. Active
435 monitoring, radical prostatectomy, or radiotherapy for localised prostate
436 cancer: Study design and diagnostic and baseline results of the ProtecT
437 randomised phase 3 trial. *Lancet Oncol*. Lane et al. Open Access article
438 distributed under the terms of CC BY; 2014;15: 1109–1118.
439 doi:10.1016/S1470-2045(14)70361-4
- 440 12. Szulkin R, Whittington T, Eklund M, Aly M, Eeles RA, Easton D, et al.
441 Prediction of individual genetic risk to prostate cancer using a polygenic
442 score. *Prostate*. 2015;75: 1467–1474. doi:10.1002/pros.23037

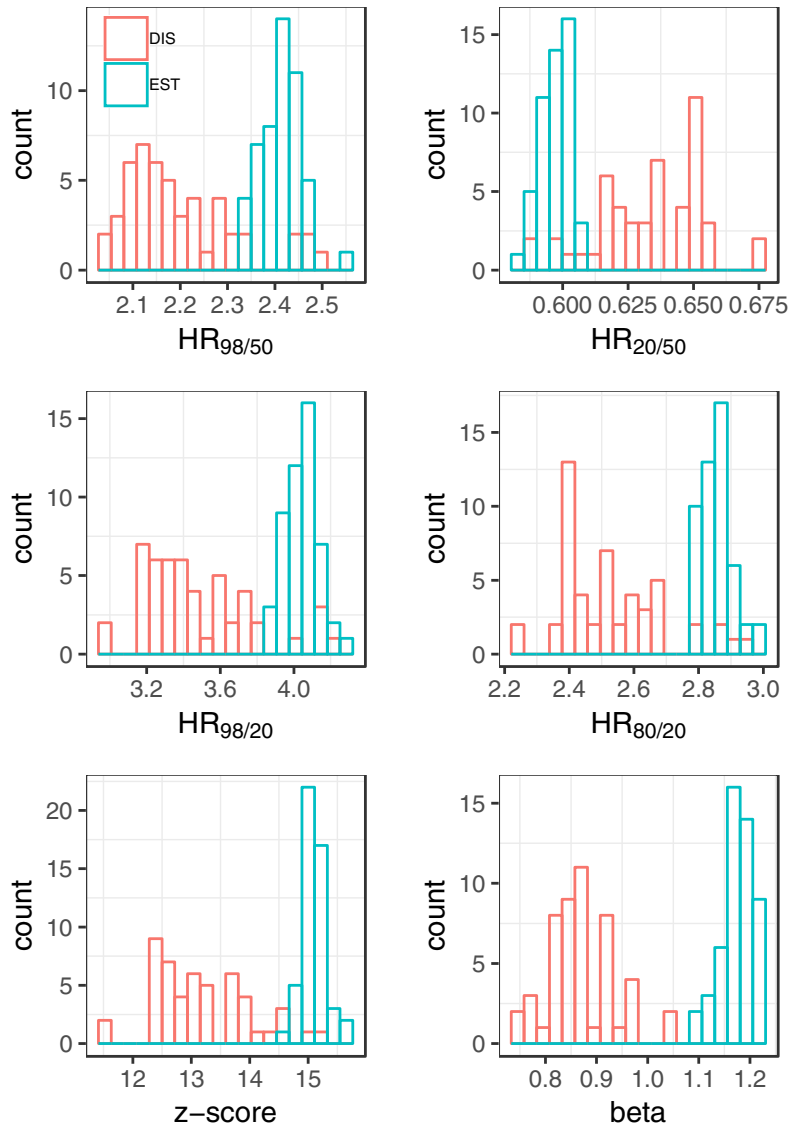
443 13. Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark

444 AG, et al. A map of human genome variation from population-scale

445 sequencing. *Nature*. 2010;467: 1061–1073. doi:10.1038/nature09534

446

447



448

449 **Figure 1.** Comparison of performance metrics between Established (EST) and Discovery (DIS)

450 SNP models using 50 random samples of the training set using a sample size of 30 thousand.

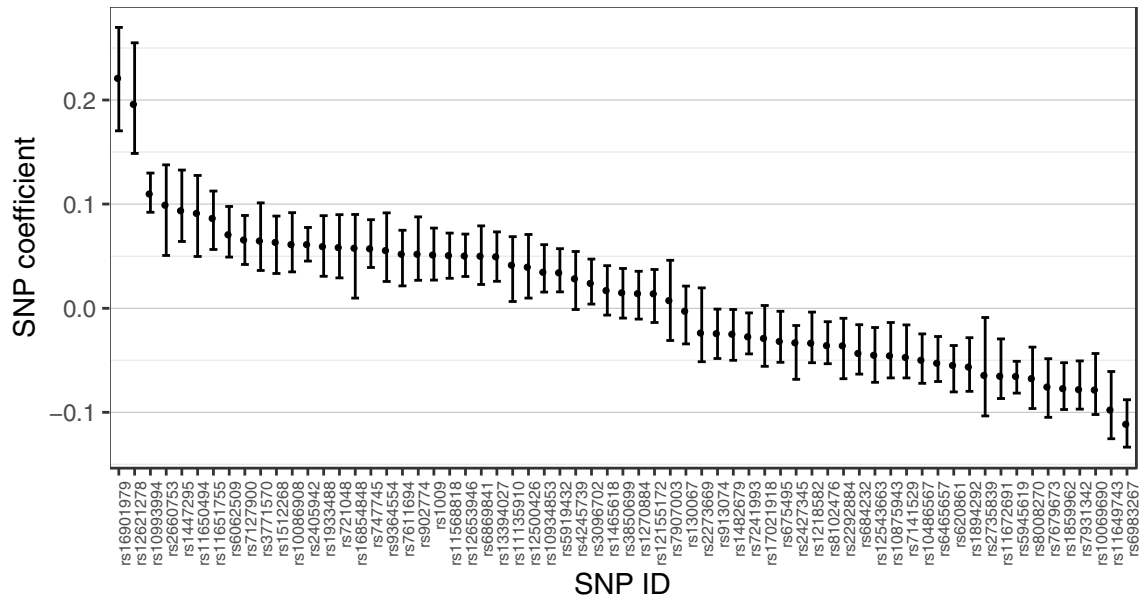
451 There is more variability with the Discovery process. Established SNPs, though, were discovered

452 using the data in the training set; this circularity is not accounted for in the present study, which

453 focuses on sample size effects.

454

455



456

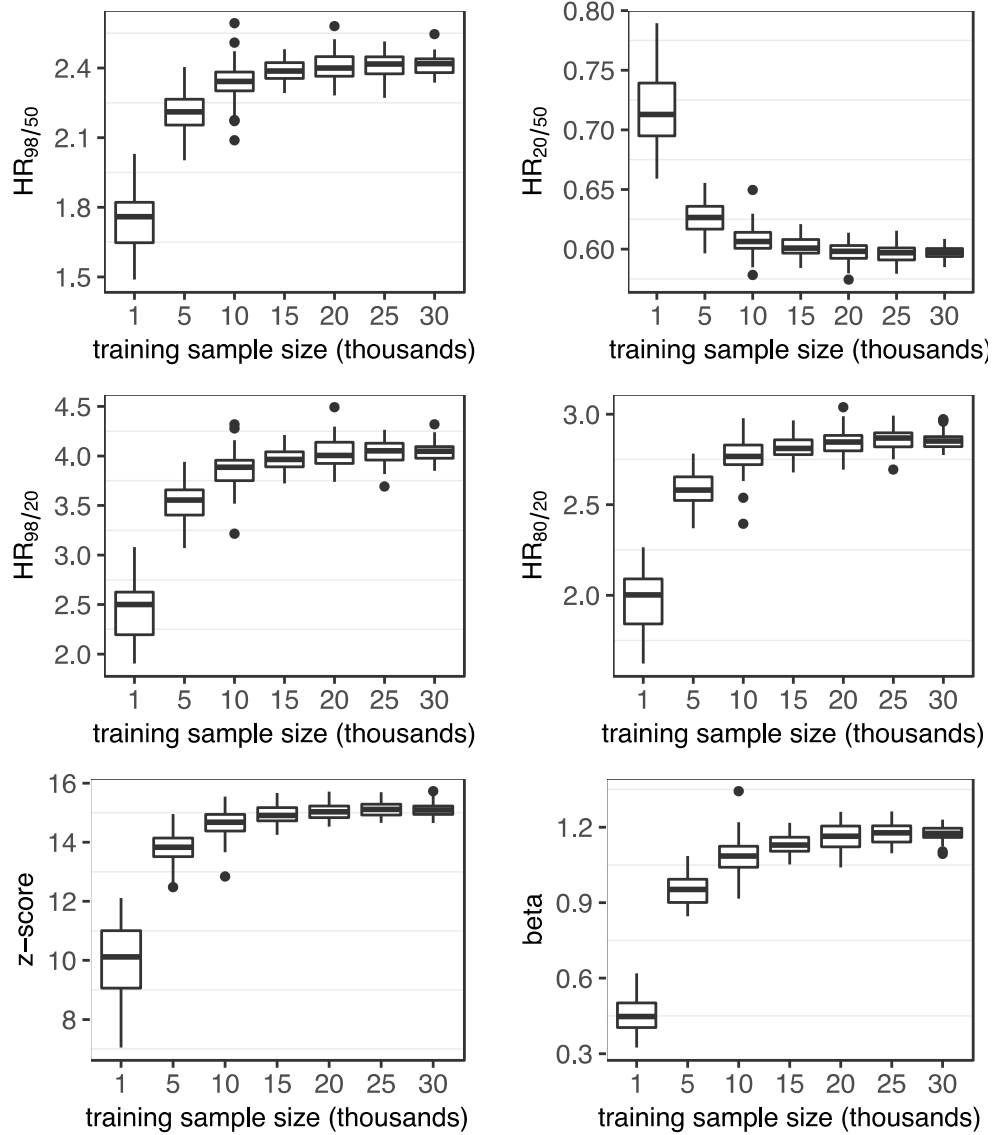
457

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.

458

459

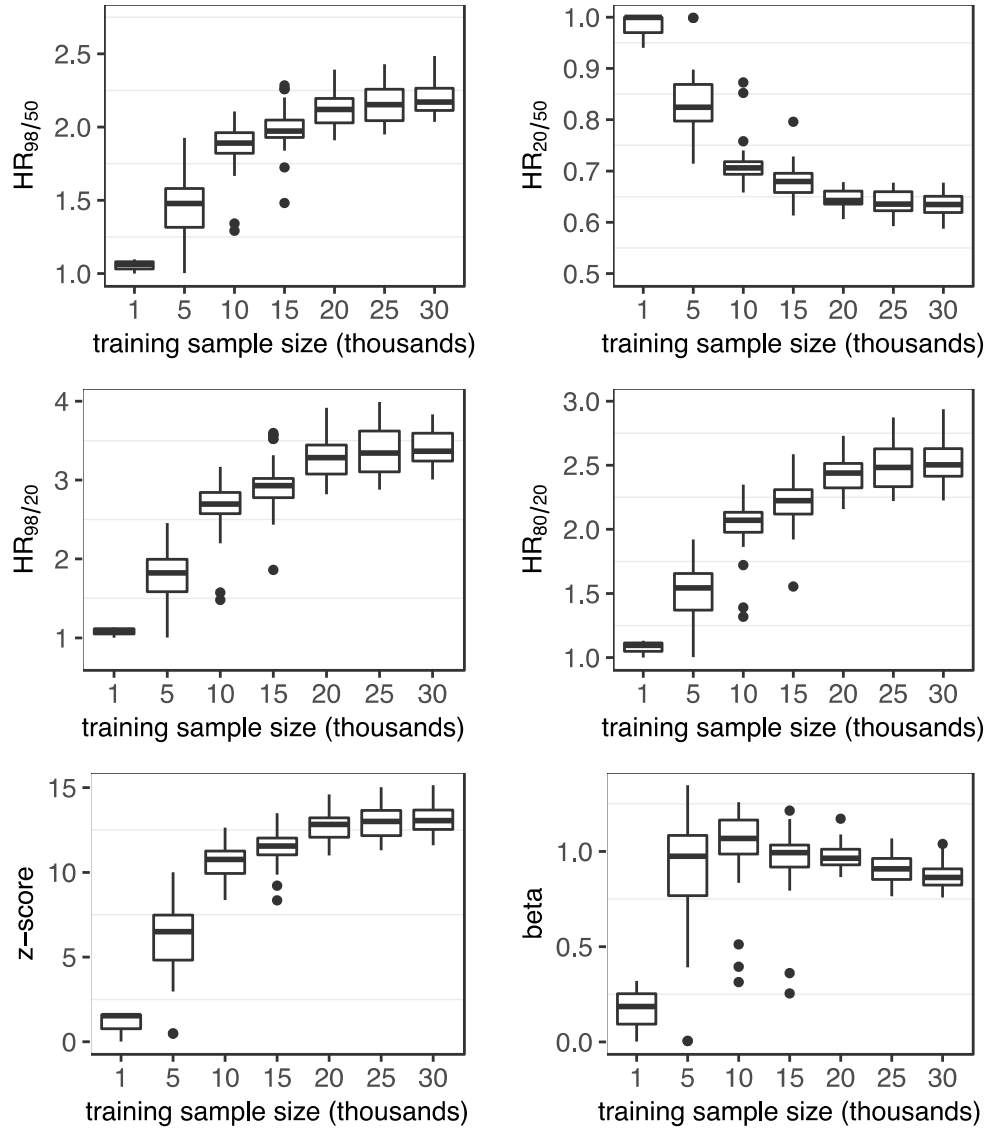
460



461

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

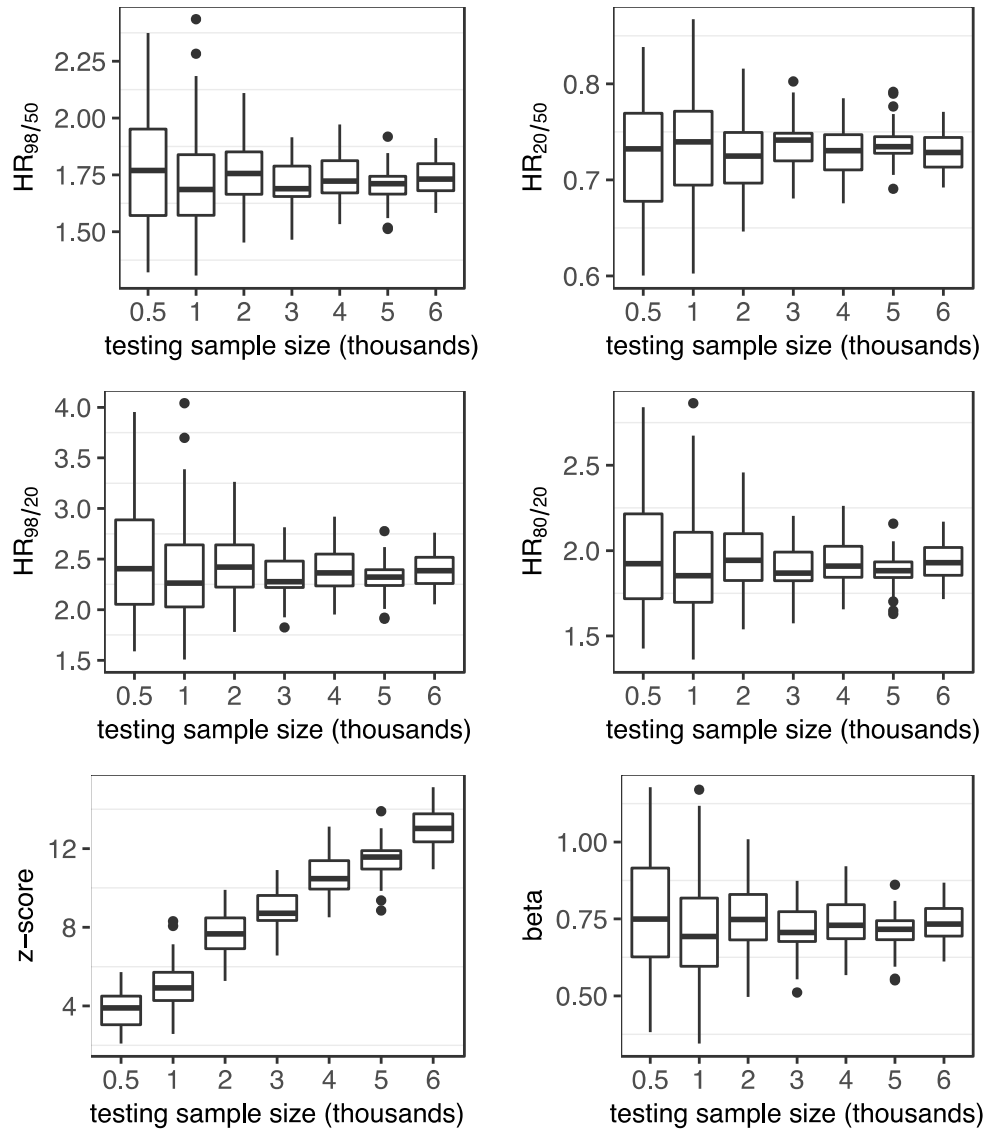
466



467

468 **Figure 4.** Performance metrics of the Discovery SNP model. Box plots of performance metrics
469 are shown for random samples of the training set using sample sizes of 1, 5, 10, 15, 20, 25, and
470 30 thousand total observations. Within each box plot, the horizontal line represents the median
471 and the box extends from the 25th to 75th percentile.

472



473

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

477

478 **Supporting Information Legends**

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