Development and validation of a novel algorithm to estimate risk of developing prostate cancer in asymptomatic men: a cohort study.

Presenting Original Research

Authors

Julia Hippisley-Cox[1]	Professor of Clinical Epidemiology & General Practice
Carol Coupland[2]	Professor of Medical Statistics in Primary Care

Institutions

[1]Nuffield Department of Primary Care Health Sciences, University of Oxford Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG, UK

[2]Division of Primary Care, 13th floor, Tower Building, University Park, University of Nottingham, Nottingham, NG2 7RD, UK

Word count 2865

Author for correspondence

Professor Julia Hippisley-Cox

Email: Julia.hippisley-cox@phc.ox.ac.uk

Telephone: +44 (0)1865 289 593

ABSTRACT

Background Early diagnosis of prostate cancer can potentially identify tumours at an early stage when intervention may help improve treatment options and survival.

Aim To develop and validate a risk prediction equation to predict absolute risk of prostate cancer in asymptomatic men with prostate specific antigen (PSA) tests in primary care.

Design Open cohort study.

Setting Routine data from 1098 QResearch® English general practices linked to mortality, hospital and cancer records for model development. Two separate sets of practices for validation.

Method 844,455 men aged 25-84 years with prostate specific antigen (PSA) tests recorded and free of prostate cancer at baseline in the derivation cohort; 292,084 and 316,583 in each validation cohort. Risk factors assessed at baseline: PSA, age, ethnicity, deprivation, BMI, smoking, family history of prostate cancer, diabetes, mental illness. Primary outcome was incident prostate cancer. Secondary outcomes were prostate cancer mortality and high-grade cancer. Cox proportional hazards models used to derive 10-year risk equations. Measures of performance were determined in both validation cohorts.

Results 40,821 incident cases of prostate cancer in the derivation cohort. The risk equation included PSA level, age, deprivation, ethnicity, smoking, family history of prostate cancer, serious mental illness, diabetes and BMI. The risk equation explained 70.4% (95%CI 69.2 to 71.6) of the variation in time to diagnosis of prostate cancer (R^2); D statistic = 3.15 (95%CI 3.06 to 3.25); Harrell's C = 0.917 (95%CI 0.915 to 0.919). The two-step approach had higher sensitivity than the fixed PSA threshold at identifying prostate cancer cases (identified 68.2% vs 43.9% of cases), high grade cancers (49.2% vs 40.3%) and deaths (67% vs 31.5%).

Conclusion

We have developed and externally validated a risk equation to quantify 10-year risk of prostate cancer in asymptomatic men undergoing a PSA test. The equation provides valid measures of absolute risk and had higher sensitivity both for incident prostate cancer, high grade cancers and prostate cancer mortality, than a simple approach based on age and PSA threshold. This warrants further validation to assess utility of the model to prioritize men in primary care for further investigation.

Web calculator

Here is a publicly available web calculator to implement the algorithm. The username/password will be removed when the paper is published. It also has the open source software for download.

Username	reviewer
Password	AfterFiveDays
URL	https://qcancer.org/10yr/prostate+psa/

How this fits in?

- Early diagnosis of prostate cancer can potentially identify tumours at an early stage when intervention may help improve treatment options and survival.
- We have developed and validate a new risk prediction equation to predict the absolute risk of prostate cancer in asymptomatic men with prostate specific antigen (PSA) tests recorded in primary care.
- The risk equation provides a valid measure of absolute risk of prostate cancer which is more efficient at identifying incident cases of prostate cancer, high grade cancers and prostate cancer deaths than an approach based on a simple PSA threshold.
- The prostate cancer risk model has the potential to prioritize patients in primary care for further investigation including imaging by multiparametric MRI.

Introduction

Prostate cancer affects an estimated one million men worldwide with almost 300,000 dying from the disease each year[1]. Prostate specific antigen (PSA) is a widely used biomarker to help detect prostate cancer before symptoms develop or at an earlier stage. Early diagnosis of prostate cancer can potentially identify tumours at an early stage when intervention may improve treatment options and survival[2]. However, multiple studies suggest the poor sensitivity of PSA alone in determining the presence of prostate cancer, for any risk stratification category[2].

A recent meta-analysis concluded that whilst screening may result in a small absolute benefit in disease specific mortality at 10 years, it does not improve overall mortality[3]. A European trial reported a 27% reduction in prostate cancer mortality attributable to PSA testing[4] at 13 years. Two other trials in the US and UK showed no overall mortality benefit[5 6] although the results might be partially explained by low adherence rates and contamination of the control group[7 8]. UK Guidelines recommend against systematic prostate cancer screening, instead allowing men aged 50 and over to request screening on demand[2]. US guidelines recommend "individualised decision making after a discussion with a clinician so each man has the opportunity to understand the potential benefits and harms of screening and incorporate his values and preferences into his decision"[9] although the tools to achieve this are largely unavailable and such shared decision making is seldom undertaken[8]. A recent BMJ rapid review which summarised all the available evidence on prostate cancer screening with PSA tests highlighted the need for research to test risk stratified approaches[8].

In other clinical areas such as , the prevention of cardiovascular disease, guidelines have evolved from clinical decisions made solely on thresholds of cholesterol, to decisions made according to absolute risk incorporating other risk factors[10 11]. As highlighted recently[8], a similar risk stratified approach could provide an effective mechanism to improve decision making for doctors and patients by providing realistic estimates of absolute risk of prostate cancer incorporating age, ethnic group, family history and other risk factors. This could also reduce unnecessary referrals since it could be applied before undertaking further investigations such as MRI or biopsies.[7 12] A systematic review identified several studies deriving risk equations for predicting absolute risk of prostate cancer incorporating PSA, although the sample sizes were small and not representative of primary care, the populations studied were predominantly white, discrimination was limited and calibration poorly reported[13]. Existing calculators have been designed to predict risk of a current diagnosis of prostate cancer rather than the future risk of developing prostate cancer and/or clinically significant disease over a 10-year period[14] [15] [16].

Currently the decision in most primary care practices to refer asymptomatic men is based on binary PSA thresholds although this can lead to too many false negative and false positive results. Furthermore, a binary threshold does not give any indication for the patient as to their absolute risk of developing prostate cancer and/or clinically significant disease requiring immediate intervention. As the diagnostic pathway has evolved considerably, at least in the UK and Europe, a PSA alone no longer triggers prostate biopsy, which is now preceded by a multiparametric MRI (mp MRI) scan. However, mp MRI misses approximately 15% of important prostate cancers and is difficult to interpret in younger men. Our aim was to develop and determine the additional predictive utility of a new algorithm to predict risk of prostate cancer for use in asymptomatic men in primary care. The intended use is to provide a better evidence base for the GP and patient to improve decision making regarding which action would be appropriate e.g. reassurance, repeat the PSA, refer for an MRI scan, regular monitoring, refer to a urologist or use of preventative interventions should any become available.

Methods

Study design, sources of data and participants

We undertook a large open cohort study of men registered with 1503 practices contributing to the QResearch® database (version 43) which is the largest and most representative GP research database in the UK[17]. We randomly allocated three quarters of practices to the derivation dataset and the remaining quarter to a validation dataset. We also identified a second validation cohort of men registered with general practices contributing to the Clinical Practice Research Data Link (CPRD-Gold)[17].

The cohorts included men aged 25-84 years registered with practices in the study period (1 January 1998 to 31st March 2018 for QResearch and 1st January 1998 to 31st March 2015 for CPRD) who had had at least one PSA level result. We excluded men with a previous diagnosis of prostate cancer at baseline and as our aim was to quantify risk in asymptomatic men we also excluded those with recorded evidence of lower urinary tract symptoms, including urinary retention, urinary frequency, nocturia, erectile dysfunction, haematuria and haematospermia in the 28 days prior to a PSA test since these were unlikely to be having PSA tests for screening purposes.

We determined an initial entry date to the cohort for each man, which was the latest of the following dates: (a) 25th birthday; (b) date of registration with the practice plus one year; (c) date on which the practice computer system was installed plus one year; (d) the beginning of the study period (01 January 1998). We then determined the date of the first PSA test during the study period after their initial entry date. This date was then used for the study entry date for the main analysis. Men were followed up until the earliest of the following dates: date of diagnosis of prostate cancer; death; de-registration with the practice; last upload of computerised data and the study end date (31st March 2018 or 31st March 2015 for CPRD). We used all the relevant patients on the database to maximise the power and also generalisability of the results.

Outcomes

Our primary outcome measure was incident diagnosis of prostate cancer during follow up as recorded on the general practice computer records or the linked hospital, mortality, or cancer registry data (where available). For mortality, we included men as having the primary outcome where prostate cancer was recorded as the main cause of death. We used the earliest recorded date of prostate cancer on any of these data sources as the outcome date. Secondary outcomes were mortality due to prostate cancer and high-grade prostate cancer as determined by the Gleason score where high-grade was a recorded combined score of 7 (4+3), 8, 9 or 10 (Gleason Grade Group 3, 4 or 5)[18].

Predictor variables

We selected variables previously found to be predictive of prostate cancer (age, selfassigned ethnicity, material deprivation (Townsend score), body mass index (BMI), smoking status, type 1 and type 2 diabetes and serious mental illness)[19] and which are recorded in patients' primary care electronic records and also PSA levels. We used the latest information recorded in the GP record on or before the study entry date (i.e. date of the first PSA test).

Derivation and validation of the models

We developed and validated a risk prediction equation for prostate cancer diagnosis using established methods[20-22]. Our initial analysis was based on patients with complete data. We then used multiple imputation with chained equations to replace missing values for BMI and smoking status for our main analyses[23-25]. We used Cox's proportional hazards models to estimate the coefficients for each predictor variable. We used Rubin's rules to combine the results across the imputed datasets[26]. We used fractional polynomials[27] to model non-linear risk relationships with continuous variables (age, BMI and PSA). We examined interactions between predictor variables and age and included significant interactions. We used the regression coefficients from the final risk equation as weights which we combined with non-parametric estimates of the baseline survivor function[28], evaluated for each year up to 15 years to derive risk equations[29]. This enabled us to derive risk estimates for each year of follow-up, with a specific focus on 10-year risk estimates.

Validation of the model

We used multiple imputation in both validation cohorts to replace missing values for BMI and smoking status. We then applied the final risk equation to both validation cohorts and calculated measures of discrimination. As in previous studies[30], we calculated D statistics[31], R^2 statistics[32] and Harrell's C statistics evaluated at 10 years. We assessed calibration by comparing the mean predicted risks at 10 years with the observed risks by tenth of predicted risk. We calculated calibration slopes. We also calculated discrimination measures for the secondary outcomes of prostate cancer mortality and high-grade cancer.

Risk stratified approach

To compare performance of the new risk prediction tool with current UK recommendations[2] we calculated the sensitivity for two different strategies for classifying men as high risk of prostate cancer (Figure 4). We then ascertained the number and proportion of all cases of diagnosed prostate cancer that would be identified over 10 years in the resulting high-risk groups (sensitivity). We also calculated the proportion of total prostate cancer deaths and the proportion of high-grade cancer cases identified by each strategy.

We used Stata (version 16) for all analyses. We adhered to the TRIPOD statement for reporting[33].

Results

Study population and incidence rates

Overall, 1457 QResearch® practices (97%) were included. Of these, 1,098 were randomly assigned to the derivation cohort with the remainder (359 practices) assigned to a validation cohort. There were 357 practices in the CPRD validation cohort. Figure 1 shows the flow of patients resulting in 844,455 men in the QResearch derivation cohort, 292,084 in the QResearch® validation cohort and 316,583 in the CPRD validation cohort.

Table 1 shows the baseline characteristics of men in the derivation and validation cohorts. In the derivation cohort, the median age was 57 years. Supplementary Table 1 shows the crude incidence rates for prostate cancer in the QResearch derivation and validation cohorts. There were 40,821 men diagnosed with prostate cancer in the QResearch derivation cohort and of these 3246 (8%) died due to prostate cancer, 11,210 (27.5%) had a high-grade Gleason score, 14,851 (36.4%) were low grade and 14,760 (36.2%) did not have a Gleason score recorded. The distribution was similar in both validation cohorts.

Predictor variables

Table 2 shows hazard ratios for men for both the complete case analysis and the multiply imputed data. The final equation included PSA, age, deprivation score, ethnicity, BMI, smoking status, family history of prostate cancer, serious mental illness and type 1 and type 2 diabetes. Increasing deprivation was associated with lower risk of prostate cancer as shown in table 2. There were significant interactions between age and family history of prostate cancer and between age and PSA levels. Supplementary figures 1a-e show graphs of the adjusted hazard ratios for the fractional polynomial terms for age, body mass index, and PSA as well as interaction terms.

Validation

The model had high levels of explained variation and discrimination in both validation cohorts (Table 3). In the QResearch validation cohort, the model explained 70.4% of the variation in time to diagnosis of prostate cancer (R^2), the D statistic was 3.15 and Harrell's C was 0.917. For prostate cancer death, the R^2 was 66.2%, the D statistic was 2.86 and Harrell's C was 0.91. For high-grade cancer, these values were 66.7%, 2.9 and 0.94 respectively. The corresponding figures in the CPRD validation cohort for prostate cancer death are shown in Table 3 (where available). Supplementary Figure 2 shows how discrimination varies across practices in the QResearch and CPRD validation cohorts.

The calibration slope was 1.03 (95% CI 1.02 to 1.04) for CPRD and 0.99 (0.98 to 1.01) for the QResearch validation cohort. Supplementary Figure 3 shows the equation is well calibrated overall and in each subgroup.

Supplementary Table 2 shows the sensitivity, specificity and observed 10-year risk based on tenths of predicted 10-year risk of prostate cancer diagnosis in the QResearch validation cohort. For example, in the top tenth of risk (i.e. men with a 10-year predicted risk of ≥20.1%), the sensitivity was 65.5%, specificity 92.6% and observed risk was 36.7%.

Risk stratification and clinical use

Figure 2 compares two strategies for identifying men at high risk of prostate cancer using the QResearch validation cohort. The two-step approach had higher sensitivity than the fixed PSA threshold at identifying prostate cancer cases (identified 68.2% vs 43.9% of cases), high grade cancers (49.2% vs 40.3%) and deaths (67% vs 31.5%).

Figure 3 shows the web calculator with clinical examples to show how the risk model could be used within a consultation. A 35-year old black Caribbean man with a PSA of 3ng/mL without a family history of prostate cancer has a 6.7% risk of prostate cancer over the next 10 years. With a family history of prostate cancer his 10-year risk of prostate cancer would be 38.2%.

Discussion

Summary

We have used the QResearch database to develop the prostate cancer risk model in asymptomatic men and externally validated it on two separate validation cohorts. Our analyses included 1.45 million men from UK primary care over a 20-year period making it substantially larger and more representative of the general population than previous studies. The results show that the risk equation provides a valid measure of absolute risk. The risk equation is more efficient at identifying incident cases of prostate cancer, high grade cancers and prostate cancer deaths than an approach based on a simple PSA threshold. We have developed a publicly available calculator to implement the algorithm which can be used to communicate levels of risk to patients to help shared decision making.

Strengths and limitations

The methods used to derive and validate these models are broadly the same as for other risk prediction equations derived from the QResearch database [34-36]. Limitations of our study include the lack of formal adjudication of diagnoses of prostate cancer although we used multiple linked data sources; potential under-ascertainment of family history of prostate cancer or high-grade Gleason scores since not all patients have recorded values. There may also be some men with undiagnosed prostate cancer in the study cohorts. Nonetheless, these limitations are likely to also occur in the clinical setting where the results are likely to be used and hence have a face validity. Key strengths include size, duration of follow up, representativeness, and lack of selection, recall and respondent bias. UK general practices have good levels of accuracy and completeness in recording clinical diagnoses and investigations and give us the ability to update the risk equation as data changes over time[37].

Comparisons with existing literature

Our hazard ratios for established predictors were similar to those reported elsewhere. Family history of prostate cancer was associated with a higher risk of prostate cancer as in other studies[38]. Black African and Caribbean men had significantly higher risks compared with white men[39]. Similarly serious mental illness was associated with a lower risk of prostate cancer as reported elsewhere[40]. Diabetes was associated with a lower risk of prostate cancer in line with previous studies[41 42] which has been postulated as being either a detection bias or a possible protective association of diabetes medication.[43]

Our study improved on the PCPT[14] and ERSPC[15 16] calculators since it was (a) developed from a large representative primary care population including almost 1 million men compared with trial populations of several thousand men already selected for biopsy; (b) it includes established risk factors; (c) it can be used to predict short and longer term absolute risks; (d) it uses existing information from electronic health records and so can be easily implemented in a primary care setting; (e) it can be updated in line with changes in the population, clinical data and clinical practice; (f) it has been externally validated; (g) the equation is published for transparency.

Implications for research and practice

We have developed and externally validated a risk equation to quantify 10-year risk of prostate cancer in asymptomatic men undergoing a PSA test. This warrants further validation to assess utility of the model to prioritize men in primary care for further investigation such as multiparametric MRI.

Other information

Acknowledgements

We acknowledge the contribution of EMIS practices who contribute to QResearch® and Egton Medical Information Systems Ltd (EMIS) and the University of Nottingham for expertise in establishing the QResearch database. This project involves data derived from patient-level information collected by the NHS, as part of the care and support of cancer patients. The data is collated, maintained and quality assured by the National Cancer Registration and Analysis Service, which is part of Public Health England (PHE). Access to the data was facilitated by the PHE Office for Data Release. The Hospital Episode Statistics data used in this analysis are re-used by permission from the NHS Digital who retain the copyright. We thank the Office of National Statistics for providing the mortality data. Public Health England, the Office of National Statistics and NHS Digital bear no responsibility for the analysis or interpretation of the data.

Ethics approval:

The QResearch® ethics approval is with East Midlands-Derby Research Ethics Committee [Reference 03/4/021 until 31.01.2019 and 18/EM/0400 from 01.02.2019].

Contributors

JHC initiated the study, developed the research question, undertook the literature review, data extraction, data manipulation and primary data analysis and wrote the first draft of the paper. CC contributed to the refinement of the research question, design, analysis, interpretation and drafting of the paper. We are grateful to Professor Freddie Hamdy (Nuffield Professor of Surgery and Professor of Urology, University of Oxford) for clinical academic input on the manuscript.

Funding

There was no external funding for this project.

Competing Interests

All authors have completed the Unified Competing Interest form at

www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: From 01.02.2019, JHC is Professor of Clinical Epidemiology and General Practice at the University of Oxford. She is founder and director of QResearch database which is not-for-profit organisation with EMIS (leading commercial supplier of IT for 55% of general practices in the UK). JHC is co-owner of ClinRisk Ltd and was a paid director until June 2019. ClinRisk Ltd develops open and closed source software to ensure the reliable and updatable implementation of clinical risk equations within clinical computer systems to help improve patient care. CC is Professor of Medical Statistics in Primary Care at the University of Nottingham and a paid consultant statistician for ClinRisk Ltd. This work and any views expressed within it are solely those of the co-authors and not of any affiliated bodies or organisations. The lead author (JHC) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Data Sharing

We have published the source code for the risk equation to ensure that those interested in reviewing the open source will find the latest available version on the website https://qcancer.org/10yr/prostate+psa/ as the equation continues to be updated. The equations presented in this paper will be released as Open Source Software under the GNU lesser GPL v3. The open source software allows use without charge under the terms of the GNU lesser public license version 3. Closed source software can be licensed at a fee.

References

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 2015;136(5):E359-86. doi: 10.1002/ijc.29210 [published Online First: 2014/09/16]
2. National Institute for Clinical Excellence. Prostate cancer: clinical knowledge summary: National Institute for Clinical Excellence,. 2017 [updated July 2017; cited 2018 31.03.2018]. Available from: https://cks.nice.org.uk/prostate-cancer#!diagnosissub:2 accessed March 2018 2018.
3. Ilic D, Djulbegovic M, Jung JH, et al. Prostate cancer screening with prostate-specific antigen (PSA) test: a systematic review and meta-analysis. *BMJ* 2018;362:k3519. doi: 10.1136/bmj.k3519
4. Schroder FH, Hugosson J, Roobol MJ, et al. Screening and prostate cancer mortality: results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. *Lancet* 2014;384(9959):2027-35. doi: 10.1016/s0140-6736(14)60525-0
5. Martin RM, Donovan JL, Turner EL, et al. Effect of a low-intensity psa-based screening intervention on prostate cancer mortality: The cap randomized clinical trial. *JAMA* 2018;319(9):883-95. doi: 10.1001/jama.2018.0154
6. Andriole GL, Crawford ED, Grubb RL, 3rd, et al. Prostate cancer screening in the randomized Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial: mortality results after 13 years of follow-up. *J Natl Cancer Inst* 2012;104(2):125-32. doi: 10.1093/jnci/djr500
7. Schroder FH. Stratifying risk--the U.S. Preventive Services Task Force and prostate-cancer screening. *N Engl J Med* 2011;365(21):1953-5. doi: 10.1056/NEJMp1112140
8. Tikkinen KAO, Dahm P, Lytvyn L, et al. Prostate cancer screening with prostate-specific antigen (PSA) test: a clinical practice guideline. *BMJ* 2018;362:k3581. doi: 10.1136/bmj.k3581 [published Online First: 2018/09/07]
9. Bibbins-Domingo K, Grossman DC, Curry SJ. The US Preventive Services Task Force 2017 Draft Recommendation Statement on Screening for Prostate Cancer: An Invitation to Review and Comment. *Jama* 2017;317(19):1949-50. doi: 10.1001/jama.2017.4413 [published Online First: 2017/04/12]
10. NHS England NE, British Medical Association. 2018/19 General Medical Services (GMS) contract QUality and Outcomes Framework (QOF): Guidance for GMS contract 2018/19, 2018:148.
11. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017;357:j2099. doi: 10.1136/bmj.j2099
12. Osses DF, Alberts AR, Bausch GCF, et al. Multivariable risk-based patient selection for prostate biopsy in a primary health care setting: referral rate and biopsy results from a urology outpatient clinic. *Transl Androl Urol* 2018;7(1):27-33. doi: 10.21037/tau.2017.12.11
13. Louie KS, Seigneurin A, Cathcart P, et al. Do prostate cancer risk models improve the predictive accuracy of PSA screening? A meta-analysis. *Ann Oncol* 2015;26(5):848-64. doi: 10.1093/annonc/mdu525
14. Thompson IM, Ankerst DP, Chi C, et al. Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial. *J Natl Cancer Inst* 2006;98(8):529-34. doi: 10.1093/jnci/djj131 [published Online First: 2006/04/20]
15. Roobol MJ, Steyerberg EW, Kranse R, et al. A risk-based strategy improves prostate-specific antigen-driven detection of prostate cancer. *European urology* 2010;57(1):79-85. doi: 10.1016/j.eururo.2009.08.025
16. Ankerst DP, Straubinger J, Selig K, et al. A Contemporary Prostate Biopsy Risk Calculator Based on Multiple Heterogeneous Cohorts. *European urology* 2018;74(2):197-203. doi: 10.1016/j.eururo.2018.05.003 [published Online First: 2018/05/21]

17. Kontopantelis E, Stevens RJ, Helms PJ, et al. Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study. *BMJ Open* 2018;8(2) doi: 10.1136/bmjopen-2017-020738

18. Epstein JI, Egevad L, Amin MB, et al. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. *The American Journal of Surgical Pathology* 2016;40(2):244-52. doi: 10.1097/pas.0000000000000530

19. Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. *BMJ Open* 2015;5(3):e007825. doi: 10.1136/bmjopen-2015-007825

20. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ* 2007:bmj.39261.471806.55. doi: 10.1136/bmj.39261.471806.55

21. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. *Heart* 2008;94:34-39. doi: 10.1136/hrt.2007.134890

22. Collins GS, Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. *BMJ* 2009;339:b2584-. doi: 10.1136/bmj.b2584

23. Schafer J, Graham J. Missing data: our view of the state of the art. *Psychological Methods* 2002;7:147-77.

24. Steyerberg EW, van Veen M. Imputation is beneficial for handling missing data in predictive models. *J Epidemiol Community Health* 2007;60:979.

25. Moons KGM, Donders RART, Stijnen T, et al. Using the outcome for imputation of missing predictor values was preferred. *J Epidemiol Community Health* 2006;59:1092.

26. Rubin DB. Multiple Imputation for Non-response in Surveys. New York: John Wiley 1987.

27. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol* 1999;28:964-74.

28. Kalbfleisch J, Prentice R. The Statistical Analysis of Failure Time Data. Hoboken2002:114-8.

29. Hosmer D, Lemeshow S, May S. Applied Survival Analysis: Regression Modelling of Time to Event data. US: Wiley 1999.

30. Hippisley-Cox J, Coupland C, Brindle P. The performance of seven QPrediction risk scores in an independent external sample of patients from general practice: a validation study. *BMJ Open* 2014;4(8):e005809. doi: 10.1136/bmjopen-2014-005809

31. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004;23:723 - 48.

32. Royston P. Explained variation for survival models. *Stata J* 2006;6:1-14.

33. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD StatementThe TRIPOD Statement. *Annals of Internal Medicine* 2015;162(1):55-63. doi: 10.7326/M14-0697

34. Hippisley-Cox J, Coupland C, Robson J, et al. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *BMJ* 2009;338:b880-. doi: 10.1136/bmj.b880

35. Hippisley-Cox J, Coupland C. Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom: prospective open cohort study. *BMJ* 2012;344(may22 1):e3427-e27. doi: 10.1136/bmj.e3427

36. Hippisley-Cox J, Coupland C. Predicting the risk of Chronic Kidney Disease in Men and Women in England and Wales: prospective derivation and external validation of the QKidney(R) Scores. *BMC Family Practice* 2010;11:49.

37. Majeed A. Sources, uses, strengths and limitations of data collected in primary care in England. *Health Statistics Quarterly* 2004(21):5-14.

38. Johns LE, Houlston RS. A systematic review and meta-analysis of familial prostate cancer risk. *BJU international* 2003;91(9):789-94.

39. Ben-Shlomo Y, Evans S, Ibrahim F, et al. The risk of prostate cancer amongst black men in the United Kingdom: the PROCESS cohort study. *European urology* 2008;53(1):99-105. doi: 10.1016/j.eururo.2007.02.047

40. Torrey EF. Prostate cancer and schizophrenia. *Urology* 2006;68(6):1280-83. doi: http://dx.doi.org/10.1016/j.urology.2006.08.1061

41. Zhang F, Yang Y, Skrip L, et al. Diabetes mellitus and risk of prostate cancer: an updated meta-analysis based on 12 case-control and 25 cohort studies. *Acta diabetologica* 2012;49 Suppl 1:S235-46. doi: 10.1007/s00592-012-0439-5 [published Online First: 2012/11/06]

42. Xu H, Mao SH, Ding GX, et al. Diabetes mellitus reduces prostate cancer risk - no function of age at diagnosis or duration of disease. *Asian Pac J Cancer Prev* 2013;14(1):441-7. [published Online First: 2013/03/29]

43. Velaer KN, Leppert JT. Diabetes Medications, Prostate-Specific Antigen Values, and the Chemoprevention of Prostate Cancer. *JAMA Netw Open* 2019;2(11):e1914644. doi: 10.1001/jamanetworkopen.2019.14644 [published Online First: 2019/11/07]

Table 1 Baseline characteristics of men aged 25-84 years free of prostate cancer and recent urinary symptoms at baseline. Values are number (%) unless indicated otherwise.

	QResearch derivation cohort	**QResearch validation cohort**	**CPRD validation cohort**
Total number of men	844,455	292,084	316,583
Median age (IQR)	57 (48-67)	57 (48-67)	58 (49-67)
Mean Townsend score (SD)	-0.5 (3.1)	-0.4 (3.1)	-1.2 (3.0)
Age band			
25-49 years	244480 (29.0)	83294 (28.5)	85087 (26.9)
50-59 years	225655 (26.7)	79029 (27.1)	89633 (28.3)
60-69 years	211355 (25.0)	72989 (25.0)	81253 (25.7)
70-84 years	162965 (19.3)	56772 (19.4)	60610 (19.1)
Ethnic group			
Ethnicity recorded	661354 (78.3)	228664 (78.3)	155947 (49.3)
White/not recorded	763692 (90.4)	264163 (90.4)	305087 (96.4)
Indian	15883 (1.9)	5428 (1.9)	2693 (0.9)
Pakistani	9501 (1.1)	3087 (1.1)	1012 (0.3)
Bangladeshi	4875 (0.6)	2003 (0.7)	254 (0.1)
Other Asian	8388 (1.0)	2642 (0.9)	1311 (0.4)
Caribbean	13198 (1.6)	4354 (1.5)	1644 (0.5)
Black African	12631 (1.5)	4704 (1.6)	1750 (0.6)
Chinese	1968 (0.2)	667 (0.2)	293 (0.1)
Other ethnic group	14319 (1.7)	5036 (1.7)	2539 (0.8)
Smoking status			
Smoking status recorded	839482 (99.4)	290479 (99.5)	314742 (99.4)
Non smoker	421809 (50.0)	144973 (49.6)	132363 (41.8)
Ex-smoker	250843 (29.7)	86556 (29.6)	78345 (24.7)
Light smoker (1-9/day)	96515 (11.4)	34647 (11.9)	51075 (16.1)
Moderate smoker (10-19/day)	36412 (4.3)	12709 (4.4)	30271 (9.6)
Heavy smoker (20+/day)	33903 (4.0)	11594 (4.0)	22688 (7.2)
median PSA (IQR)	1.18 (1.82)	1.16 (1.76)	1.22 (2.09)
BMI recorded	672319 (79.6)	234612 (80.3)	237333 (75.0)
mean BMI kg/m^2 (SD)	27.2 (4.4)	27.2 (4.4)	26.7 (4.0)
Family history of prostate cancer	8881 (1.1)	2884 (1.0)	1999 (0.6)
Serious mental illness	6475 (0.8)	2386 (0.8)	1946 (0.6)
Type 1 diabetes	2652 (0.3)	890 (0.3)	849 (0.3)
Type 2 diabetes	65406 (7.7)	23070 (7.9)	18512 (5.8)

Table 2: Adjusted hazard ratios (95% CI) for prostate cancer diagnosis for the complete case analysis (n=661,354) and analysis based on multiply imputed data (n=844,455 with 5 imputed datasets). For fractional polynomial terms & age interactions see footnotes and figures.

	Unadjusted HR (95% CI) Complete case analysis	**Adjusted HR (95% CI) Complete case analysis**	**Adjusted HR (95% CI) Imputed data**
Deprivation Score (5-unit increase)Ψ	0.83 (0.82 to 0.85)	0.91 (0.89 to 0.93)	0.91 (0.90 to 0.93)
Ethnic Group			
White or not recorded	1.00	1.00	1.00
Indian	0.40 (0.36 to 0.45)	0.67 (0.59 to 0.75)	0.67 (0.60 to 0.75)
Pakistani	0.29 (0.24 to 0.35)	0.54 (0.45 to 0.64)	0.54 (0.46 to 0.65)
Bangladeshi	0.16 (0.12 to 0.23)	0.46 (0.33 to 0.65)	0.47 (0.33 to 0.66)
Other Asian	0.33 (0.29 to 0.40)	0.59 (0.50 to 0.71)	0.60 (0.50 to 0.72)
Black Caribbean	1.54 (1.44 to 1.65)	1.56 (1.46 to 1.67)	1.56 (1.46 to 1.67)
Black African	0.80 (0.73 to 0.88)	1.13 (1.02 to 1.25)	1.14 (1.04 to 1.26)
Chinese	0.50 (0.37 to 0.67)	0.54 (0.40 to 0.72)	0.55 (0.41 to 0.73)
Other ethnic group	0.74 (0.68 to 0.82)	1.09 (0.99 to 1.20)	1.10 (1.00 to 1.21)
Smoking status			
Non smoker	1.00	1.00	1.00
Ex-smoker	1.02 (1.00 to 1.05)	1.01 (0.98 to 1.03)	1.00 (0.98 to 1.03)
Light smoker	0.86 (0.83 to 0.89)	0.97 (0.94 to 1.01)	0.98 (0.95 to 1.02)
Moderate smoker	0.77 (0.72 to 0.82)	0.93 (0.88 to 0.99)	0.93 (0.88 to 0.99)
Heavy smoker	0.74 (0.69 to 0.79)	0.94 (0.88 to 1.00)	0.95 (0.90 to 1.01)
Family history of prostate cancer†	1.47 (1.34 to 1.61)	1.73 (1.55 to 1.92)	1.83 (1.66 to 2.02)
Serious mental Illness‡	0.52 (0.44 to 0.63)	0.67 (0.56 to 0.80)	0.67 (0.57 to 0.79)
No diabetes	1.00	1.00	1.00
Type 1 diabetes	0.35 (0.25 to 0.49)	0.74 (0.53 to 1.04)	0.78 (0.58 to 1.05)
Type 2 diabetes	0.78 (0.75 to 0.82)	0.90 (0.86 to 0.95)	0.90 (0.86 to 0.94)

†interaction with age; hazard ratio evaluated at mean age

‡compared with patients without this characteristic

Ψ increasing levels of Townsend score indicate increasing levels of deprivation

Model also includes fractional polynomial terms for age (age$^{-0.5}$, age$^{-0.5}$ln(age)) and BMI (BMI^{-1}, BMI $^{-0.5}$) and PSA (PSA^{-1}, PSA$^{-0.5}$) with interaction terms between age terms and family history and between age and PSA terms

Table 3 Performance of the risk model to predict prostate cancer diagnosis, prostate cancer death and high-grade prostate cancer in the QResearch validation and CPRD validation cohorts, comparing complete case and imputed datasets

	QResearch Validation Cohort (complete case n=188,013 patients)	**QResearch Validation Cohort (imputed data n=292,084 patients)**	**CPRD Validation Cohort (complete data n=120,869 patients)**	**CPRD Validation Cohort (imputed data n=316,583 patients)**
	Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CI)
Prostate cancer diagnosis				
Harrell's C	0.920 (0.917 to 0.923)	0.917 (0.915 to 0.919)	0.922 (0.919 to 0.925)	0.916 (0.914 to 0.918)
D statistic	2.71 (2.67 to 2.75)	3.15 (3.06 to 3.25)	2.83 (2.78 to 2.87)	2.82 (2.79 to 2.85)
R^2 Statistic	63.7 (62.8 to 64.5)	70.4 (69.2 to 71.6)	65.6 (64.5 to 66.7)	65.5 (65.1 to 65.9)
Prostate cancer death				
Harrell's C	0.909 (0.895 to 0.923)	0.907 (0.897 to 0.917)	0.901 (0.865 to 0.937)	0.906 (0.894 to 0.918)
D statistic	2.84 (2.69 to 2.99)	2.86 (2.76 to 2.97)	3.10 (2.78 to 3.42)	3.16 (3.04 to 3.28)
R^2 Statistic	65.9 (63.4 to 68.3)	66.2 (64.6 to 67.8)	69.6 (64.8 to 73.6)	69.4 (68.8 to 72.0)
High grade prostate cancer				
Harrell's C	0.934 (0.930 to 0.939)	0.935 (0.932 to 0.938)	n/a	n/a
D statistic	2.88 (2.82 to 2.95)	2.90 (2.85 to 2.95)	n/a	n/a
R^2 Statistic	66.5 (65.3 to 67.7)	66.7 (65.9 to 67.6)	n/a	n/a