Application of a novel machine learning framework for predicting non-metastatic prostate cancer-specific mortality in men using the Surveillance, Epidemiology, and End Results (SEER) database

Changhee Lee, Alexander Light*, Ahmed Alaa, David Thurtle, Mihaela van der Schaar†, Vincent J Gnanapragasam†*

Summary

Background Accurate prognostication is crucial in treatment decisions made for men diagnosed with non-metastatic prostate cancer. Current models rely on prespecified variables, which limits their performance. We aimed to investigate a novel machine learning approach to develop an improved prognostic model for predicting 10-year prostate cancer-specific mortality and compare its performance with existing validated models.

Methods We derived and tested a machine learning-based model using Survival Quilts, an algorithm that automatically selects and tunes ensembles of survival models using clinicopathological variables. Our study involved a US population-based cohort of 171 942 men diagnosed with non-metastatic prostate cancer between Jan 1, 2000, and Dec 31, 2016, from the prospectively maintained Surveillance, Epidemiology, and End Results (SEER) Program. The primary outcome was prediction of 10-year prostate cancer-specific mortality. Model discrimination was assessed using the concordance index (c-index), and calibration was assessed using Brier scores. The Survival Quilts model was compared with nine other prognostic models in clinical use, and decision curve analysis was done.

Findings 647 151 men with prostate cancer were enrolled into the SEER database, of whom 171 942 were included in this study. Discrimination improved with greater granularity, and multivariable models outperformed tier-based models. The Survival Quilts model showed good discrimination (c-index 0·829, 95% CI 0·820–0·838) for 10-year prostate cancer-specific mortality, which was similar to the top-ranked multivariable models: PREDICT Prostate (0·820, 0·811–0·829) and Memorial Sloan Kettering Cancer Center (MSKCC) nomogram (0·787, 0·776–0·798). All three multivariable models showed good calibration with low Brier scores (Survival Quilts 0·036, 95% CI 0·035–0·037; PREDICT Prostate 0·036, 0·035–0·037; MSKCC 0·037, 0·035–0·039). Of the tier-based systems, the Cancer of the Prostate Risk Assessment model (c-index 0·782, 95% CI 0·771–0·793) and Cambridge Prognostic Groups model (0·779, 0·767–0·791) showed higher discrimination for predicting 10-year prostate cancer-specific mortality. c-indices for models from the National Comprehensive Cancer Care Network, Genitourinary Radiation Oncologists of Canada, American Urological Association, European Association of Urology, and National Institute for Health and Care Excellence ranged from 0·711 (0·701–0·721) to 0·761 (0·750–0·772). Discrimination for the Survival Quilts model was maintained when stratified by age and ethnicity. Decision curve analysis showed an incremental net benefit from the Survival Quilts model compared with the MSKCC and PREDICT Prostate models currently used in practice.

Interpretation A novel machine learning-based approach produced a prognostic model, Survival Quilts, with discrimination for 10-year prostate cancer-specific mortality similar to the top-ranked prognostic models, using only standard clinicopathological variables. Future integration of additional data will likely improve model performance and accuracy for personalised prognostics.

Funding None.

Lancet Digit Health 2021; 3: e158–65

Published Online February 3, 2021 https://doi.org/10.1016/S2589-7500(20)30314-9

See Comment page e138

*Joint first authors

†Joint senior author

Department of Electrical and Computer Engineering, University of California, Los Angeles, CA, USA (C Lee MS, A Alaa PhD); **Department of Surgery, Division of Urology** (A Light MBBS, D Thurtle MRCS, V J Gnanapragasam FRCS[Urol]), **Department of Applied Mathematics and Theoretical Physics** (Prof M van der Schaar PhD), **Department of Public Health and Primary Care** (Prof M van der Schaar), **and Cambridge Urology Translational Research and Clinical Trials Office, Cambridge Biomedical Campus** (V J Gnanapragasam), **University of Cambridge, Cambridge, UK; Department of Urology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK** (A Light, D Thurtle, V J Gnanapragasam)

Correspondence to: Dr Vincent J Gnanapragasam, Cambridge Urology Translational Research and Clinical Trials Office, Cambridge Biomedical Campus, University of Cambridge, Cambridge CB2 0QQ, UK **vjg29@cam.ac.uk**

Introduction

Prostate cancer is the most common cancer in men and its global incidence is rising.[1] More than 80% of men diagnosed present with non-metastatic disease. Treatment decisions are particularly complex and need to balance the risk of progression with therapy-related morbidity.[2] Accurate prognostication is therefore crucial to identify patients who could benefit most from treatment.[3,4]

Many nationally and internationally endorsed tools for risk modelling are available. Most stratify men into risk groups and are derived from the three-tiered D'Amico system, originally developed to predict biochemical

Research in context

Evidence before this study

Prognostic models for non-metastatic prostate cancer have been built using traditional statistical modelling with prespecified variables and interactions. These typically place patients into risk groups or categories using clinicopathological variables. However, a major aim for future health care is to make treatment decisions more personalised, particularly for men diagnosed with non-metastatic prostate cancer, for whom treatment choices and decisions are complex. Machine learning systems offer the possibility of individualising predictions for this population, but no such tools are in use. We searched PubMed from database inception to April 10, 2020, using the search term "prostate cancer artificial intelligence". This search identified very few machine learning studies in prostate cancer prognostics, mostly small, single ethnic cohorts, and proof-of-concept studies. There are no studies in large population cohorts, and to our knowledge, none have compared model performance with or added value to currently used risk prediction models.

Added value of this study

We used a large (n=171 942), multiethnic, population-based, prospectively maintained SEER dataset to produce a model trained using machine learning to predict 10-year prostate cancer-specific mortality. We used a novel algorithm called Survival Quilts, which exploits an ensemble of traditional and machine learning-based modelling techniques. The survival function learned by Survival Quilts is a combination of survival profiles from these modelling techniques; therefore, the algorithm is optimised to account for discriminative performance and calibration. The Survival Quilts model produced in this study predicted 10-year prostate cancer-specific mortality with good discrimination and was well calibrated. In comparison with nine other models in clinical use, our model showed similar discrimination to the top-ranked models in predicting the outcome, even when stratified by age and ethnicity. Further, we observed that applying Survival Quilts in clinical decision model analysis might be beneficial. To our knowledge, this is the first study that shows the advantages of a data-driven, variable-agnostic, machine learning approach in predicting prostate cancer-specific mortality. This approach will probably improve with further training on new datasets and the addition of variables (eg, new imaging or molecular markers). With development and refinement, this method could be used clinically to provide superior, more individualised, survival predictions.

Implications of all the available evidence

Clinicians and patients need to balance the risks and benefits of treatment and consider multiple variables that might affect prognosis. Machine learning algorithms can adapt to quickly integrate data from multiple variables, such as those in prostate cancer, for individual prognostic modelling. The data-driven and variable-agnostic approach inherent to machine learning also allows for an information gain from previously unsuspected contributing factors. Therefore, machine learning could form the basis for a new era of prognostic models that more accurately predict individualised survival outcomes and enhance decision making in prostate cancer and other cancers.

recurrence.[5–10] However, biochemical recurrence is a poor surrogate for survival, and prognostic models should therefore be based on survival outcomes.[11,12] The simple combination of prostate-specific antigen (PSA), grade, and stage can enable the creation of effective prognostic models, and refining group-stratification systems can improve model discrimination.[5,9,10,13] Additionally, studies[14–16] have shown that using continuous data rather than categorisation can make prognostication more accurate and personalised for clinical decision making. For example, the PREDICT Prostate tool[14] and Memorial Sloan Kettering Cancer Center (MSKCC) nomogram[15] have shown high discriminative ability for predicting survival in robust external validation and are available as accessible web-based decision aids for patients and clinicians.

However, even these more personalised tools rely on traditional statistical modelling, with prespecified variables and interactions. Machine learning is a data-driven application of artificial intelligence, in which systems automatically learn and improve without explicit programming. Accordingly, machine learning is able to autonomously exploit datasets to identify new variables and more complex relationships between them. Its application is growing rapidly in health care and is increasingly being used to develop novel prognostic models in several diseases.[17] We hypothesise that machine learning might produce a superior predictive model for prostate cancer. For prostate cancer prognostication, machine learning has so far been restricted to small, proof-of-concept studies without comparison with reference standards.[18–21] We aimed to investigate a novel machine learning approach, Survival Quilts, using a large national observational dataset to develop an improved prognostic model for predicting 10-year prostate cancer-specific mortality in men with non-metastatic disease,[22] and compare its performance with existing validated models available in clinical practice.

Methods

Data source and study population

We used data collected from the Surveillance, Epidemiology, and End Results (SEER) Program. SEER collects data regarding cancer diagnoses and survival for approximately 30% of the US population, and benefits from extensive quality review.[23] Men aged 35–95 years

diagnosed with histologically confirmed non-metastatic prostate cancer (site code C61·9) between Jan 1, 2000, and Dec 31, 2016, were included. Men with evidence of metastatic disease (including lymph node metastasis); those with missing survival data or data on PSA, Gleason grade, or stage; and men younger than 35 years or older than 95 years were excluded. Complete data were required for PSA, Gleason grade, stage, and prostate cancer-specific mortality. Time-to-event or censoring was derived from the date of diagnosis or the date of last contact (either death or last follow-up). Biopsy core involvement was available for 66 885 (38·9%) of 171 942 men in the final cohort and derived by mean imputation where missing. Access to the SEER database does not need formal ethics approval and is covered by its open access policy.

Model development

The following variables, measured at diagnosis, were included in model development: age, PSA, primary and secondary Gleason grades or grade groups, T stage, total number of biopsy cores examined, and core positivity (number of cores positive for cancer divided by number of cores taken). MRI, comorbidity, and treatment data were not available. We derived our machine learning-based survival model using Survival Quilts: an open-source software developed to automate deployment of machine learning in survival analysis.[22] Survival Quilts is an ensemble of different survival models. Because different models exhibit varying discriminative performance and calibration accuracy between datasets, Survival Quilts learns to automatically weigh these models and tune the parameters of each individual model in a single ensemble for the specific dataset. The survival function learned by Survival Quilts is a combination of survival profiles produced by many models, optimised to account for discriminative performance and calibration. Therefore, Survival Quilts is a superset of many statistical models and machine learning-based models for survival prediction. This software is automated and frees researchers from choosing one survival model without the need for in-depth knowledge of machine learning. The four models included in this study ranged from traditional statistical models to state-of-the-art deep learning models: Cox proportional hazards, random survival forest, conditional inference survival forest, and DeepHit models (appendix p 1).[24–26] The turning parameters were chosen using a grid search, on the basis of validation performance determined by the concordance index (c-index), to predict prostate cancer-specific mortality.[22] The primary outcome was prediction of prostate cancer-specific mortality at 10 years, assessed in men aged 35–95 years with non-metastatic prostate cancer. The SEER cohort was randomly split (64:16:20) into the training, validation, or testing sets, using the Python (version 3.6.5) package scikit-learn. For model evaluation, we used bootstrapping of 10 000 patients in the testing set, with more than 100 iterations on average. Time-dependent c-indices were calculated for model discrimination and Brier scores for calibration. Model calibration, reflecting predicted versus observed outcomes, was also assessed by visual inspection of calibration plots. Discrimination was assessed in the full cohort and then stratified by age groups on the basis of the cohort median, resulting in the following age groups: younger than 65 years (n=79 003) and 65 years or older (n=92 939). We also stratified by different ethnicities (Black, White, other; appendix p 4).

Head-to-head model comparison

The Survival Quilts model was compared with nine other prognostic models in clinical use. The attributes of each model are shown in the appendix (p 2). The following models were included: the tier-based Cambridge Prognostic Groups, European Association of Urology, National Institute for Health and Care Excellence (NICE), Genitourinary Radiation Oncologists of Canada, American Urological Association, and the National Comprehensive Cancer Care Network.[5–10] We also compared Survival Quilts with the point-based Cancer of the Prostate Risk Assessment model, and the multivariable MSKCC nomogram and PREDICT Prostate models.[14–16] Due to unavailable data on treatment and comorbidity for the PREDICT Prostate model, we removed these variables from hazard calculations. Using the testing set, model performance at 10 years was compared by calculating the c-index to show how well models discriminate the risk of prostate cancer-specific mortality. A sensitivity analysis was also done without the biopsy core involvement variable. For each of the three models, the probability of 10-year prostate cancer-specific mortality risk was calculated across a range of threshold probabilities and plotted against default strategies of treating all or no patients regardless of prognosis. Decision curve analysis was used to calculate a clinical net benefit for each prediction model. The net benefit of following these strategies was compared with the three models (Survival Quilts, PREDICT Prostate, and MSKCC) for a prognosis-based intervention—ie, an intervention with a predicted risk that exceeds a specific risk threshold. Net benefit was defined as the value achieved by making decisions on the basis of model predictions. The statistical tools used for these analyses included R (version 3.6.1) and Python (version 3.6.5), and relevant packages are detailed in the appendix (p 3).

For the **SEER database** see https://seer.cancer.gov/data/access.html

See **Online** for appendix

Role of the funding source

There was no funding source for this study.

Results

Figure 1 shows our data assembly process. 647 151 men were enrolled into the SEER database with histologically confirmed prostate cancer in the study period between Jan 1, 2000, and Dec 31, 2016. Of these, 7340 did not have survival data or censoring information, 21 528 presented with evidence of lymph nodes or

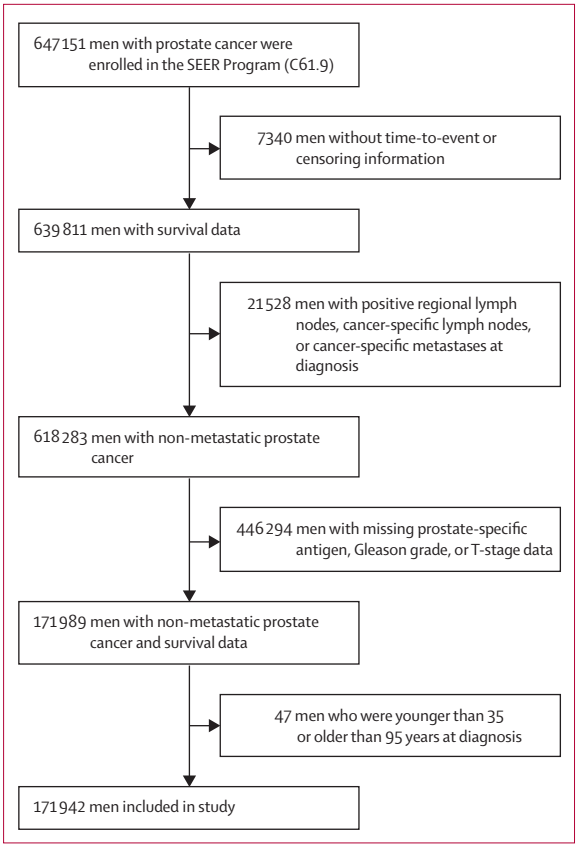

Figure 1: **Patient data selection process**
SEER=Surveillance, Epidemiology, and End Results.

	Overall (n=171 942)
Mean age, years	65·6 (8·9)
Mean prostate-specific antigen, ng/mL	10·1 (13·3)
Mean core involvement*	
Cores taken	12·4 (2·5)
Cores positive	4·2 (2·1)
Cores negative	8·1 (2·7)
T stage	
1a	1811 (1·1%)
1b	1026 (0·6%)
1c	101 036 (58·8%)
2a	48 690 (28·3%)
2b	11 282 (6·6%)
2c	4728 (2·8%)
3a	1699 (1·0%)
3b	1195 (0·7%)
4	475 (0·3%)
Primary Gleason score	
1	0
2	9 (0·0%)
3	126 083 (73·3%)
4	42 588 (24·8%)
5	3262 (1·9%)
Secondary Gleason score	
1	0
2	6 (0·0%)
3	94 715 (55·1%)
4	67 284 (39·1%)
5	9937 (5·7%)
Grade group	
1	72 548 (42·2%)
2	52 245 (30·4%)
3	21 806 (12·7%)
4	14 675 (8·5%)
5	10 668 (6·2%)
Prostate cancer-specific mortality	
5 years	2469 (1·4%)
10 years	4157 (2·4%)
All-cause mortality	
5 years	14 825 (8·6%)
10 years	26 488 (15·4%)

Data are mean (SD) or n (%). *Data available for 66 885 (38·9%) of 171 942 men.

Table 1: **Demographic characteristics of patients included in the analysis**

metastasis, and 446 294 had missing data for at least one of the essential domains of PSA, Gleason grade, or clinical stage data. 47 men were outside of the study age range (35–95 years) and were also excluded. The final study population included 171 942 men. Table 1 shows baseline characteristics of the study cohort, and ethnicities are shown in the appendix (p 4). The mean age was 65·6 years (SD 8·9). Most men were White (78·0% [134 139] of 171 942), 14·2% (24 488) were Black, and 5·2% (8962) were Asian. Almost all cancers were T stage 1 or 2 (168 573 [98·0%] of 171 942) and grade group 1–3 (146 599 [85·3%]). Only a low proportion of patients (0–0·01%) had a primary Gleason score of less than 3 and were included. Median time to event for men who died with prostate cancer was 4·4 years (IQR 4·5) and for the remaining cohort (including other causes of death) was 6·2 years (IQR 5·7), giving an overall median follow-up of 6·1 years. At 10 years, 4157 of 171 942 men died of prostate cancer, and 26 488 died of other causes.

The c-index for predicting prostate cancer-specific mortality was consistently high in training, validation, and testing sets (0·829, 95% CI 0·820–0·838), with good calibration (Brier score 0·036, 0·035–0·037; table 2; figure 2). c-index was also high when the cohort was subdivided by age (table 3) For Survival Quilts, model performance was marginally better in men younger than 65 than in men aged 65 years or older (table 3). Next, we tested performance in different ethnic groups. The Survival Quilts model was consistent, with small differences in the c-index between ethnicities (table 4).

Our Survival Quilts model was favourable in terms of performance, compared with other tier-based and multivariable models. Variables used for these different modelling systems are shown in the appendix (p 2). The Cancer of the Prostate Risk Assessment model

	c-index (95% CI)	Brier score (95% CI)
Multivariable models		
Survival Quilts	0·829 (0·820–0·838)	0·036 (0·035–0·037)
PREDICT Prostate	0·820 (0·811–0·829)	0·036 (0·035–0·037)
Memorial Sloan Kettering Cancer Center	0·787 (0·776–0·798)	0·037 (0·035–0·039)
Tier-based models		
Cancer of the Prostate Risk Assessment score	0·782 (0·771–0·793)	0·037 (0·035–0·039)
Cambridge Prognostic Groups	0·779 (0·767–0·791)	0·037 (0·035–0·039)
National Comprehensive Cancer Care Network	0·761 (0·750–0·772)	0·038 (0·036–0·040)
Genitourinary Radiation Oncologists of Canada	0·750 (0·739–0·761)	0·039 (0·037–0·041)
American Urological Association	0·749 (0·738–0·760)	0·039 (0·037–0·041)
European Association of Urology	0·711 (0·701–0·721)	0·039 (0·037–0·041)
National Institute for Health and Care Excellence	0·711 (0·701–0·721)	0·039 (0·037–0·041)

Data are shown for the testing set. c-index=concordance index.

Table 2: **Discrimination and calibration of each model at predicting 10-year prostate cancer-specific mortality**

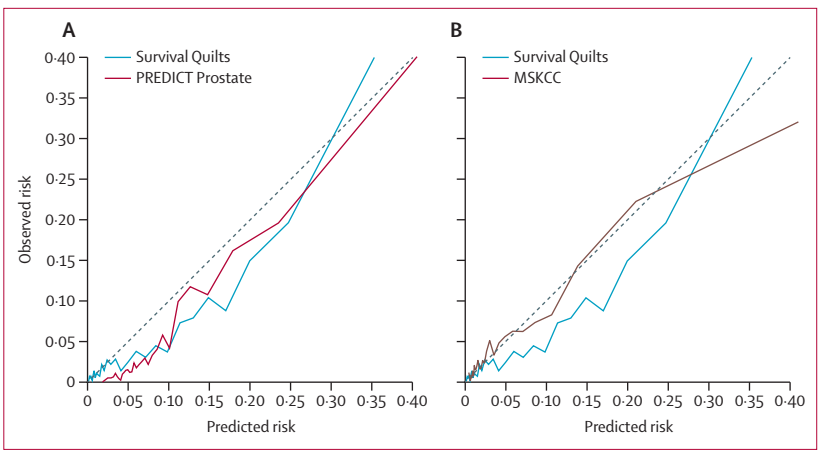

Figure 2: **Calibration plots of observed versus predicted risk**
Prostate cancer-specific mortality at 10 years, assessed in men aged 35–95 years with non-metastatic prostate cancer. Survival Quilts model compared with the top two performing prognostic models: (A) PREDICT Prostate and (B) MSKCC model. MSKCC=Memorial Sloan Kettering Cancer Center.

	Age <65 years (n=79 003)	Age ≥65 years (n=92 939)
Survival Quilts	0·834 (0·817–0·851)	0·797 (0·786–0·808)
PREDICT Prostate	0·819 (0·802–0·836)	0·789 (0·778–0·800)
Memorial Sloan Kettering Cancer Center	0·830 (0·813–0·847)	0·749 (0·737–0·761)
Cancer of the Prostate Risk Assessment score	0·818 (0·801–0·835)	0·742 (0·730–0·754)
Cambridge Prognostic Groups	0·824 (0·807–0·841)	0·742 (0·729–0·755)
National Comprehensive Cancer Care Network	0·807 (0·790–0·824)	0·725 (0·713–0·737)

c-index=concordance index.

Table 3: **Comparative c-index for 10-year prostate cancer-specific mortality stratified by age**

(c-index 0·782, 95% CI 0·771–0·793) and Cambridge Prognostic Groups model (0·779, 0·767–0·791) showed higher discrimination for predicting 10-year prostate cancer-specific mortality (table 2). c-indices for models from the National Comprehensive Cancer Care Network, Genitourinary Radiation Oncologists of Canada, American Urological Association, European Association of Urology, and NICE ranged from 0·711 (0·701–0·721) to 0·761 (0·750–0·772). The multivariable models generally discriminated 10-year prostate cancer-specific mortality similarly to the tier-based models, with a c-index of 0·820 (95% CI 0·811–0·829) for the PREDICT Prostate model and 0·787 (0·776–0·798) for the MSKCC model. Our Survival Quilts model had a similarly high c-index in this cohort (0·829, 0·820–0·838). Model discrimination was maintained when the cohort was stratified by age (table 3) and ethnicity (table 4).

All three multivariable models showed good calibration with low Brier scores (Survival Quilts 0·036, 95% CI 0·035–0·037; PREDICT Prostate 0·036, 0·035–0·037; MSKCC 0·037, 0·035–0·039; table 2; figure 2). We further tested whether these comparisons were valid given that the PREDICT Prostate and MSKCC models were originally derived from different cohorts. We refitted these models to the training set, before reapplying them to the validation set. We found similar performance characteristics for the PREDICT Prostate model and improved c-index performance for the MSKCC model (appendix p 5). Both models performed similarly to the Survival Quilts model. Finally, given that biopsy core data were only available in less than half of the cohort (66 885 [38·9%] of 171 942 men), we reassessed the PREDICT Prostate and Survival Quilts models without this variable and found similar performance characteristics (appendix p 6).

We next assessed model performance using decision curve analysis, considering the effect on treatment decision making (eg, surveillance *vs* radical treatment). The heterogeneous profile of the patient population renders a uniform treatment strategy (treat all or no patients) inferior to strategies informed by any one of the three models (figure 2). Among the three multivariable models, MSKCC provided the least net benefit whereas Survival Quilts provided the greatest gain. The gain from Survival Quilts was particularly higher than the PREDICT Prostate model with threshold probabilities of risk between 0·1 and 0·3 (figure 3), with added net incremental benefits across each threshold. The difference was even greater when compared with the MSKCC model (figure 3; appendix pp 7–8).

Discussion

In this study we used a large dataset to develop and test a machine learning-trained prognostic model for predicting 10-year prostate cancer-specific mortality and assessed its performance against a range of tiered and multivariable prediction models. To our knowledge, our study is the

	White (n=134 139)	Black (n=24 488)	Other* (n=13 315)
Survival Quilts	0·832 (0·824–0·841)	0·815 (0·795–0·834)	0·836 (0·813–0·860)
PREDICT Prostate	0·825 (0·816–0·833)	0·802 (0·783–0·822)	0·827 (0·802–0·851)
Memorial Sloan Kettering Cancer Center	0·779 (0·767–0·790)	0·811 (0·791–0·831)	0·778 (0·745–0·811)
Cancer of the Prostate Risk Assessment score	0·773 (0·762–0·785)	0·801 (0·781–0·821)	0·775 (0·744–0·807)
Cambridge Prognostic Groups	0·774 (0·762–0·787)	0·799 (0·777–0·820)	0·744 (0·707–0·780)
National Comprehensive Cancer Care Network	0·759 (0·747–0·771)	0·777 (0·755–0·798)	0·728 (0·693–0·764)

*Full list of ethnicities shown in the appendix (p 4). c-index=concordance index.

Table 4: **Comparative c-index for 10-year prostate cancer-specific mortality stratified by ethnicity**

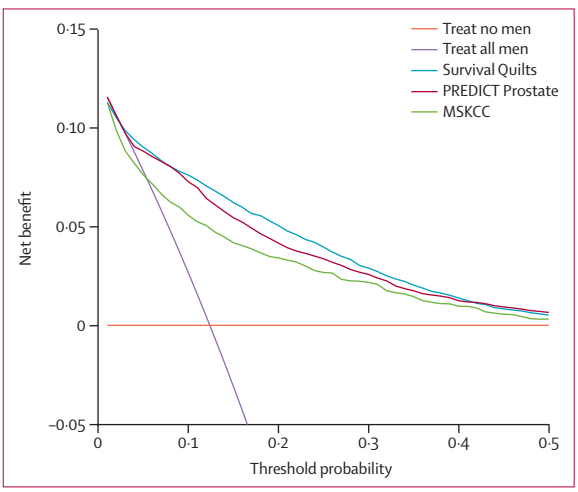

Figure 3: **Decision curve analysis**
The clinical net benefit for each prediction model is calculated across a range of risk threshold probabilities. Clinical net benefit is defined as the minimum probability of disease at which further intervention would be warranted. MSKCC=Memorial Sloan Kettering Cancer Center.

first to use the SEER cohort to compare numerous models for predicting prostate cancer-specific mortality. We observed that multivariable models outperform tiered systems, which is consistent with previous findings of head to head comparisons.[13,27] Our study introduces an innovative approach to predicting mortality by using a novel machine learning algorithm that automatically combines optimal attributes from different modelling methods.

There are few machine learning studies in prostate cancer prognostics, and these studies have included small cohorts in terms of model development. The only study[18] to assess prostate cancer-specific mortality trained several artificial neural network models with 19 pretreatment variables in 7267 Korean men. Koo and colleagues[18] long short-term memory artificial neural network model produced a c-index of 0·815 for discrimination of 10-year prostate cancer-specific mortality and all-cause mortality. However, the model has not yet been externally validated. The largest study[19] to date using machine learning in prostate cancer prognostics included data from 8581 Taiwanese men. A support vector machine-trained model incorporated comorbidity data with standard clinicopathological variables (but not PSA) and showed an accuracy of 0·852 at predicting cancer-related post-treatment recurrence and mortality.

Datasets are often limited by which variables prostate cancer specialists have traditionally considered important and therefore collected (eg, clinicopathological and comorbidity variables). Our study supports the notion that model performance improves with greater granularity, and machine learning-trained models should have an advantage when incorporating new variables.[13,18–21,28] As an example, Donovan and colleagues[20] combined standard variables with five molecular biomarkers and automated histopathological image analysis to derive a prediction tool for biochemical recurrence after treatment. Their Precise Post-op model had a c-index of 0·77 for recurrence-free survival. Zhang and colleagues[21] combined somatic mutation signatures in a 43 gene panel with NICE risk criteria and improved the area under the curve for prediction of postsurgical biochemical recurrence from 0·62 to 0·75.

The data-driven and variable-agnostic approach inherent to machine learning allows for an information gain from previously unsuspected contributing factors. For example, the machine learning AutoPrognosis model for predicting cardiovascular risk was trained on 473 variables and identified walking pace as the third most important variable for death after systolic blood pressure and body-mass index.[17] The Survival Quilts approach used in our study also permits a modelling gain whereby the most robust model among several can be objectively chosen without previous assumptions regarding model characteristics and variable interactions.[22] Notably, by using just a few standard clinicopathological factors (rather than employing new or different variables) the method was able to achieve high c-indices and good calibration.

In the decision curve analysis, we also found an incremental gain in net benefit when Survival Quilts was applied compared with the other top two performing models. There is no consensus on what is a clinically useful range for a net benefit in treatment prognostic models.[29] In clinical practice, if the uncertainty is 10% or less then a decision model is not needed. Therefore, we reasoned that threshold probabilities higher than 10% would benefit from using a decision model. In this analysis we found a net benefit when the threshold probabilities of risk were between 0·1 and 0·3. We accept that there might be other interpretations of a clinically important range, but believe that our approach was pragmatic to define a range for which prognostic model improvements have a clinical benefit on decision making. Training the Survival Quilts model on multiple large datasets and incorporating more factors will probably produce a more superior model than the one we achieved in this study. Owing to the autonomous nature

of machine learning, such models could be quickly and automatically updated whenever new data become available. A further opportunity is the ability to continuously add data as the patient's treatment progresses and visualise the effects on prognosis, which is not currently possible with static prognostic models.

Our machine learning-based model was trained in a large, contemporary, ethnically heterogeneous population using real-world data from a high-quality database.[24] To our knowledge, this study is the largest to apply machine learning to prostate cancer prognostics. However, this study has some important limitations. The cohort distribution is heavily skewed to earlier stage disease and represents a population heavily screened using PSA. Consequently, few death events occurred, and follow-up was limited. However, the performance of other prognostic models we tested was consistent with results seen in large population studies that had a more balanced case-mix.[13,27] We did not have data for a larger starting population and cannot account for bias that might have been introduced when the final cohort was derived. We also did not explore the geographical distribution of this US cohort or any social differences, so we cannot comment on the effects that these might have had on our results. Similarly, biopsy core data had to be imputed for a substantial proportion of the cohort and we acknowledge that using imputation for such a large amount of information might have introduced bias. As the SEER database does not collect comorbidity or treatment data, we could not model the effect of comorbidity on outcomes nor could we consider the effects of treatment, both of which are key variables in other models (eg, PREDICT Prostate). Comparison of different models could also introduce inherent bias because of input variable heterogeneity, and changes in performance might reflect this heterogeneity. SEER also does not collect data on prostate MRI, although whether MRI findings improve prognostic capabilities remains unclear.[30] We additionally did not have any molecular markers to assess, although their addition to standard models does show some promise.[31] Questions remain on the usefulness of these tests and how their addition to models such as Survival Quilts, PREDICT Prostate, or the MKSCC nomogram will improve performance given the substantial additional cost. This study focused on non-metastatic cancer, but in future work we would be keen to take the methods here and apply them to metastatic cancer, for which there is paucity of robust and validated models.

Our novel machine learning-trained model is capable of predicting 10-year prostate cancer-specific mortality at similar performance to the top existing models. Machine learning might confer numerous future advantages, especially its potential to readily incorporate new data, its ability to self-train, and its flexibility for input variables to evolve. Consequently, machine learning represents a unique future framework for producing more granular, individualised, and iterative prognostic models. This study showed in a PSA-screened population the crucial need to move away from tier-based risk grouping and increasingly use multivariable and more personalised prognostic models to guide patient management.

Contributors

CL contributed to the conceptualisation, data curation and verification, formal analysis, investigation, methods, project administration, validation, and visualisation. AL and DT contributed to the visualisation and writing the original draft. AA contributed to the conceptualisation, methods, data verification, supervision, and validation. MvdS contributed to the conceptualisation, methods, project administration, and supervision. VJG contributed to the conceptualisation, supervision, visualisation, and writing the original draft. All authors contributed to writing the Article and editing and have approved the final manuscript. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication. CL and AA have accessed and verified all the data in the study.

Declaration of interests

We declare no competing interests.

Data sharing

Data for this study were made available to us through application to the SEER database. Our current agreement precludes open sharing and forwarding of data. Access to these data can be requested through the SEER website. The source code for the analysis is available at https://bitbucket.org/mvdschaar/mlforhealthlabpub/src/master/alg/survivalquilts. All authors had access to the raw data and CL and AA verified the data. There was no funder organisation involved in data access, curation, or verification.

Acknowledgments

We acknowledge infrastructure support from the National Institute for Health Research Cambridge Biomedical Research Centre. We acknowledge the report on the use of PSA from SEER https://ascopost.com/issues/june-25–2015/national-cancer-institute-pulls-psa-data-from-seer/ released by the National Cancer Institute. We have done this analysis on the assumption that the SEER data received were corrected and updated. This study is not intended to yield clinical tools for implementation and is meant only for research purposes. Any findings need further validation in external datasets. We note that applying prognostic models developed in other datasets to the SEER dataset have shown similar prognostic performance, which provides some reassurance as to the use of PSA in our analysis.

References

1 Taitt HE. Global trends and prostate cancer: a review of incidence, detection, and mortality as influenced by race, ethnicity, and geographic location. *Am J Men Health* 2018; **12:** 807–23.
2 Donovan JL, Hamdy FC, Lane JA, et al. Patient-reported outcomes after monitoring, surgery, or radiotherapy for prostate cancer. *N Engl J Med* 2016; **375:** 1425–37.
3 Hamdy FC, Donovan JL, Lane JA, et al. 10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. *N Engl J Med* 2016; **375:** 1415–24.
4 Wilt TJ, Jones KM, Barry MJ, et al. Follow-up of prostatectomy versus observation for early prostate cancer. *N Engl J Med* 2017; **377:** 132–42.
5 Gnanapragasam VJ, Lophatananon A, Wright KA, Muir KR, Gavin A, Greenberg DC. Improving clinical risk stratification at diagnosis in primary prostate cancer: a prognostic modelling study. *PLoS Med* 2016; **13:** e1002063.
6 Mottet N, Bellmunt J, Bolla M, et al. EAU-ESTRO-SIOG guidelines on prostate cancer. Part 1: screening, diagnosis, and local treatment with curative intent. *Eur Urol* 2017; **71:** 618–29.
7 Graham J, Kirkbride P, Cann K, Hasler E, Prettyjohns M. Prostate cancer: summary of updated NICE guidance. *BMJ* 2014; **348:** f7524.
8 Lukka H, Warde P, Pickles T, Morton G, Brundage M, Souhami L. Controversies in prostate cancer radiotherapy: consensus development. *Can J Urol* 2001; **8:** 1314–22.
9 Sanda MG, Cadeddu JA, Kirkby E, et al. Clinically localized prostate cancer: AUA/ASTRO/SUO guideline. Part I: risk stratification, shared decision making, and care options. *J Urol* 2018; **199:** 683–90.

10 Mohler JL, Armstrong AJ, Bahnson RR, et al. Prostate cancer, version 1.2016. *J Natl Compr Canc Netw* 2016; **14:** 19–30.

11 Jhaveri FM, Zippe CD, Klein EA, Kupelian PA. Biochemical failure does not predict overall survival after radical prostatectomy for localized prostate cancer: 10-year results. *Urology* 1999; **54:** 884–90.

12 Kattan MW, Hess KR, Amin MB, et al. American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine. *CA Cancer J Clin* 2016; **66:** 370–74.

13 Zelic R, Garmo H, Zugna D, et al. Predicting prostate cancer death with different pretreatment risk stratification tools: a head-to-head comparison in a nationwide cohort study. *Eur Urol* 2020; **77:** 180–88.

14 Thurtle DR, Greenberg DC, Lee LS, Huang HH, Pharoah PD, Gnanapragasam VJ. Individual prognosis at diagnosis in nonmetastatic prostate cancer: Development and external validation of the PREDICT Prostate multivariable model. *PLoS Med* 2019; **16:** 1002758.

15 Stephenson AJ, Kattan MW, Eastham JA, et al. Prostate cancer-specific mortality after radical prostatectomy for patients treated in the prostate-specific antigen era. *J Clin Oncol* 2009; **27:** 4300–05.

16 Cooperberg MR, Broering JM, Carroll PR. Risk assessment for prostate cancer metastasis and mortality at the time of diagnosis. *J Natl Cancer Inst* 2009; **101:** 878–87.

17 Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. *PLoS One* 2019; **14:** e0213653.

18 Koo KC, Lee KS, Kim S, et al. Long short-term memory artificial neural network model for prediction of prostate cancer survival outcomes according to initial treatment strategy: development of an online decision-making support system. *World J Urol* 2020; **38:** 2469–76.

19 Lin YT, Lee MT, Huang YC, Liu CK, Li YT, Chen M. Prediction of recurrence-associated death from localized prostate cancer with a Charlson comorbidity index-reinforced machine learning model. *Open Med (Wars)* 2019; **14:** 593–606.

20 Donovan MJ, Fernandez G, Scott R, et al. Development and validation of a novel automated Gleason grade and molecular profile that define a highly predictive prostate cancer progression algorithm-based test. *Prostate Cancer Prostatic Dis* 2018; **21:** 594–603.

21 Zhang S, Xu Y, Hui X, et al. Improvement in prediction of prostate cancer prognosis with somatic mutational signatures. *J Cancer* 2017; **8:** 3261–67.

22 Lee C, Zame WR, Alla AM, van der Schaar M. Temporal quilting for survival analysis. *Proc Mach Learn Res* 2019; **89:** 596–605.

23 National Cancer Institute Surveillance, Epidemiology, and End Results Prostate-Specific Antigen Working Group, Adamo MP, Boten JA, et al. Validation of prostate-specific antigen laboratory values recorded in Surveillance, Epidemiology, and End Results registries. *Cancer* 2017; **123:** 697–703.

24 Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat* 2008; **2:** 841–60.

25 Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *J Comput Graph Stat* 2006; **15:** 651–74.

26 Lee C, Zame WR, Yoon J, van der Schaar M. Dynamic-DeepHit: a deep learning approach for dynamic survival analysis with competing Risks based on longitudinal data. 2019. https://par.nsf.gov/servlets/purl/10099761 (accessed April 18, 2020).

27 Thurtle D, Bratt O, Stattin P, Pharoah P, Gnanapragasam V. Comparative performance and external validation of the multivariable PREDICT prostate tool for non-metastatic prostate cancer: a study in 69,206 men from Prostate Cancer data Base Sweden (PCBaSe). *BMC Med* 2020; **18:** 139.

28 Thurtle D, Rossi SH, Berry B, Pharoah P, Gnanapragasam VJ. Models predicting survival to guide treatment decision-making in newly diagnosed primary non-metastatic prostate cancer: a systematic review. *BMJ Open* 2019; **9:** e029149.

29 Van Calster B, Wynants L, Verbeek JFM, et al. Reporting and interpreting decision curve analysis: a guide for investigators. *Eur Urol* 2018; **74:** 796–804.

30 Vickers A, Carlsson SV, Cooperberg M. Routine use of magnetic resonance imaging for early detection of prostate cancer is not justified by the clinical trial evidence. *Eur Urol* 2020; **78:** 304–06.

31 Spratt DE, Zhang J, Santiago-Jiménez M, et al. Development and validation of a novel integrated clinical-genomic risk group classification for localized prostate cancer. *J Clin Oncol* 2018; **36:** 581–90.