Survival Analysis of Localized Prostate Cancer with Deep Learning

Xin Dai[1], Ji Hwan Park[1, 2], Shinjae Yoo[1], Nicholas D'Imperio[1], Benjamin H. McMahon[3], Christopher T. Rentsch[4, 6, 7], Janet P. Tate[4, 5], and Amy C. Justice[4, 5]

[1]Computational Science Initiative, Brookhaven National Laboratory, Upton, NY, USA
[2]School of Computer Science, The University of Oklahoma, Norman, OK, USA
[3]Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM, USA
[4]VA Connecticut Healthcare System, West Haven, CT, USA
[5]Schools of Medicine and Public Health, Yale University, New Haven, CT, USA
[6]Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA
[7]Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK

ABSTRACT

In recent years, data-driven, deep learning-based models have shown great promise in medical risk prediction. By leveraging the large-scale electronic health record (EHR) data in the Department of Veterans Affairs (VA), the largest integrated healthcare system in the United States, we have developed an automated, personalized risk prediction model to support the clinical decision- making process for localized prostate cancer patients. Our method combines the representative power of deep learning and the analytical interpretability of parametric regression models. By utilizing different neural network structures, we can either handle time-dependent or static input. To have a comprehensive evaluation of model performances, we calculate time-dependent C-statistics C_{td} over short (2-yr), mid (5-yr), and long (10-yr) time horizons using either a composite outcome or prostate cancer mortality as the target event. The composite outcome combines the Prostate-Specific Antigen (PSA) test, metastasis, and prostate cancer mortality. Our longitudinal model RDSM achieved C_{td} 0.85 (0.85), 0.80 (0.84), 0.76 (0.81), and the cross-sectional model DSM got C_{td} 0.85 (0.82), 0.80 (0.82), 0.76 (0.79) for 2-yr, 5-yr and 10-yr for the composite (mortality) outcome, respectively. In addition to estimating an individual's survival probability, our models can also quantify the uncertainty associated with the prediction. The uncertainty scores showed a consistent correlation with the prediction accuracy. We found PSA and prostate cancer stage information were the most important indicators in the risk prediction. Our work demonstrates the utility of the data-driven machine learning model in prostate cancer risk prediction, which can play a critical role in the clinical decision system.

Introduction

Prostate cancer is one of the most prevalent cancers among men in the United States. Approximately 12.5% of men will be diagnosed with prostate cancer during their lifetime[1]. In the United States, aggressive prostate cancer screening leads to early diagnosis and medical intervention. However, treatment can incur severe side effects, e.g., incontinence, erectile dysfunction[2,3]. It is critical to balance the trade-offs between different treatment options to maximize the quality of life and minimize unnecessary side effects in those unlikely to benefit from treatment. Localized disease (T1-4, N0, M0)accounts for 74.3% of prostate cancer diagnosis, and 5-yr survival approaches 100%. On the other hand, prostate cancer still accounts for 5.6% of all cancer-related deaths. A comprehensive risk estimation model for prostate cancer patients will be of great clinical value.

So far, there have been several related studies utilizing different datasets and survival analysis methods [4–8] for prostate cancer. The rapid advancement in machine learning techniques, in particular, deep learning (DL), has made it possible to develop a personalized and automated risk prediction model to assist clinical decision making[9]. One of the most significant obstacles in developing such a model is the lack of large-scale and high-quality electronic health record (EHR) data. As the largest integrated healthcare system in the United States, VA has collected more than 20 years of EHR data on 30 million veterans from multiple regions into the VA Corporate Data Warehouse (CDW). Via the collaboration between the Department of Veterans Affairs (VA) and the Department of Energy (DOE), we have access to EHR and cancer registry data on more than 100, 000 veterans diagnosed with localized prostate cancer. Our goal in this study is to take advantage of the large-scale, longitudinal, national EHR data from the VA and high-performance computing power from the DOE to develop a risk prediction model for localized prostate cancer patients using cutting-edge DL methods.

Results

Following the patient selection protocol outlined in Fig. 1, our study cohort comprised 112,276 localized prostate cancer patients, 7663 (3126) patients had the composite (mortality) outcomes before censoring. The median age at the time of diagnosis was 65.5 yrs. And most of the patients were either white (66%) or black (28%) (Fig. 2). Table 1 lists the clinical feature distributions of our entire cohort.

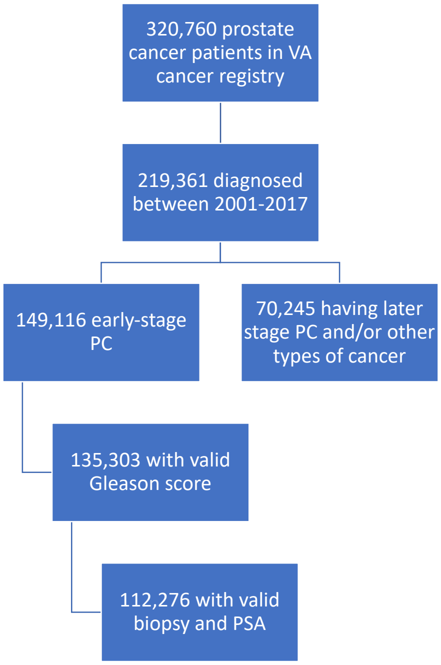
320,760 prostate cancer patients in VA cancer registry
219,361 diagnosed between 2001-2017
149,116 early-stage PC
70,245 having later stage PC and/or other types of cancer
135,303 with valid Gleason score
112,276 with valid biopsy and PSA

Figure 1. Patient selection flow chart. After gathering all prostate cancer patients diagnosed between 2001-2017 in the VA cancer registry, we excluded patients satisfying any of the following criteria: 1) late-stage (metastasized) prostate cancer at the time of diagnosis. 2) No valid PSA test 1yr before diagnosis. 3) No valid Gleason score (Unknown or < 6). 4) No biopsy record for diagnosis. 5) Having other types of cancer.

We randomly split all the patients into training (80%) and test set (20%) while ensuring the censoring ratio is consistent. Table 2 shows detailed outcome statistics of training and test sets. Since patients may experience multiple events before censoring, the number of composite outcomes is less than the sum of three single outcomes.

Table 3 summarizes model performances on the test set at different event horizons of two outcomes. For deep learning models DSM and RDSM, we used 15% of the randomly selected training data as the validation set and performed hyper-parameter optimization. The results reported in Table 3 were obtained from the test set using the models having the lowest loss on the validation set. We adopted a similar strategy for optimizing the two traditional machine learning models, GBM and RSF, and selected model parameters with the highest C_{td} on the validation set. Whereas for the Cox model, we performed 5-fold cross-validation on the entire training set and reported the result on the test set using the parameter set having the highest average C_{td} in the cross-validation. To maximize the performance, we performed separate hyper-parameter tuning against two outcomes for each model.

From Table 3, we see that across different time horizons and target outcomes, our proposed deep learning models consistently outperformed other machine learning models and the Cox model. The cross-sectional model DSM

Mean PSA(ng/ml)	8.83
Mean age at diagnosis	65.80
PSA counts	7.26
T-Stage 1	70.28%
T-Stage 2	27.96%
T-Stage 3	1.59%
T-Stage 4	0.17%
Gleason Score 6	41.29%
Gleason Score 7	42.32%
Gleason Score 8	9.55%
Gleason Score 9	6.31%
Gleason Score 10	0.53%

Table 1. Clinical feature distributions of the cohort. Here the mean PSA refers the value of the last PSA test prior to diagnosis. PSA counts is the number of PSA tests up to 10-yrs before diagnosis. The Gleason score is the sum of primary and secondary score.

Outcome	PSA > 50 ng/ ml	Metastasis	PC mortality	Composite Outcome	Right-censored
Training set	4199	2410	2494	6130	83690 (93.17%)
Test set	1062	605	632	1533	20923 (93.17%)

Table 2. Event statistics of different outcomes in training and test set.

performed the best for the composite outcome. While for the mortality outcome, the longitudinal model RDSM got a significantly higher C_{td} than the rest of the models.

Uncertainty Quantification

The significance of uncertainty quantification (UQ) is that it yields a meaningful metric about how confident the model is regarding the prediction. As a bonus of our deep learning approach (See Methods for details), we were able to calculate $v(t \mid X)$, the standard deviation of $S(t \mid X)$ for each prediction. From Table 4 we found that for both RDSM and DSM, the correlations between the Brier scores and $v(t \mid X)$ were consistent, i.e., a lower variance indicated a lower Brier score. This result suggests that $v(t \mid X)$ is a reliable indicator for UQ, which is of great importance in real-world clinical decision-making.

Subgroup Analysis

To study how our model perform in different race and age subgroups, we conducted subgroup analysis using the same train-test protocol as in the full cohort case. Table 5 and 6 show the example results for age subgroup (65-75 yrs) and race subgroup (black), respectively. Results of other subgroups can be found in the supplemental materials.

Compared with Table 3, we see that all models show deteriorating performances across different age groups, especially for the long-term (10-yr) prediction. A plausible reason is the age specification in each subgroup leads to the reduced variance in the outcome, which in turn could have a negative impact on the C_{td}. Another explanation, which is more relevant to deep learning models, is the shrinkage of patient numbers in each subgroup (Fig. 2). Empirically, deep learning models are susceptible to sample size reduction as they have more parameters to fit than traditional machine learning and statistical models. But even for the two largest age groups (55-65 and 65-75 yrs, Table 13 and 5), where the impact of sample size was less severe, the performance gaps were still significant.

Another interesting observation was if we focus on the 10-yr time horizon, the C-index for the composite outcome was relatively stable, while for the PC-mortality, the C-index first decreased with age, then increased for the oldest age group (>75 yrs).

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	0.850	**0.797**	0.755
DSM	**0.853**	0.796	**0.758**
GBM	0.849	0.795	0.753
RSF	0.850	0.790	0.746
Cox	0.837	0.787	0.743

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	**0.851**	**0.836**	**0.805**
DSM	0.821	0.824	0.789
GBM	0.799	0.820	0.786
RSF	0.801	0.818	0.779
Cox	0.805	0.817	0.780

Table 3. C_{td} for all tested models at near (2-yr), mid (5-yr) and long (10-yr) time horizons. The left (right) table shows the C_{td} for the composite (mortality) outcome. RDSM (DSM) is designed to handle time-dependent (-independent) data. See Methods section for details.

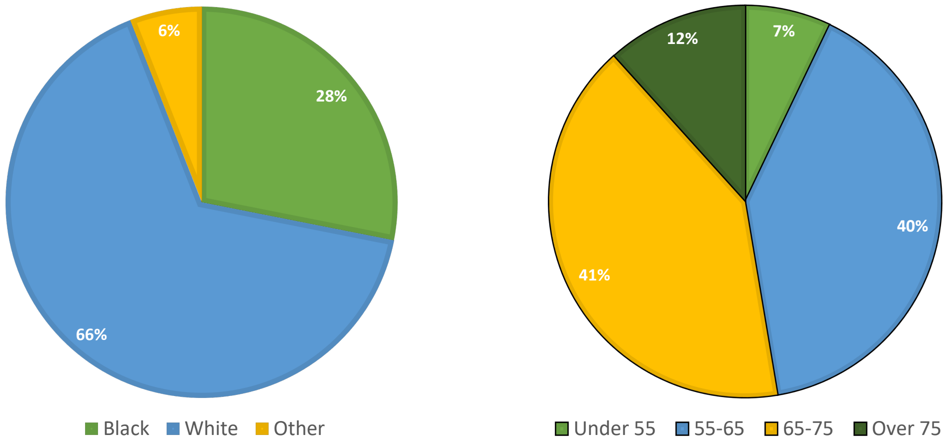

Figure 2. Distribution of race and age at diagnosis. The "Other" in the race includes all the Asian, Pacific Islander, Native American people, and patients without valid race information.

On the other hand, C_{td} remained more stable in different race groups, at least for the longer-term predictions. For example, RDSM achieved C_{td} 0.84 and 0.78 (0.83), 0.78 (0.78) for 5 and 10-yr PC-mortality prediction in the Black subgroup (Table 6), while for the White subgroup the corresponding numbers are 0.83 and 0.78, respectively (Table 15). The insensitivity of C_{td} regarding race indicates that race is not a useful indicator in predicting the PC prognosis for our models.

Ablation Study

Although our deep learning-based models got higher C_{td} and lower Brier scores compared to benchmark models, the black-box nature of deep learning hindered the interpretability. To alleviate the problem, we conducted an ablation study to quantify how each input feature contributes to the risk estimation of DSM and RDSM.

We divided input features into four categories: PSA, age, race, and cancer stages. For the longitudinal model, we further included the time-interval information. Our ablation approach was to drop each feature group and track model performances. Though bear in mind that neural networks in our deep learning models involve non-linear interactions between each input, we should remain cautious in interpreting the results.

Table 7 shows the ablation results for composite and PC-mortality outcomes. We found that race was the least important feature for both RDSM and DSM. And the second least important feature was the age at diagnosis, especially for the composite outcome. This result was consistent with the results of the subgroup analysis. We also found that the cancer stage information (Gleason score and T-stage) was crucial for the model performances, especially in the longer-term horizon (5 and 10-yr). Both RDSM and DSM experienced huge performance drops without PSA-related features for the composite outcome. Naively, it was due to our definition of the composite

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	$0.017 \mid 1.6 * 10^{-3}$	$0.040 \mid 3.8 * 10^{-3}$	$0.076 \mid 6.7 * 10^{-3}$
DSM	$0.017 \mid 2.3 * 10^{-3}$	$0.033 \mid 4.0 * 10^{-3}$	$0.060 \mid 6.5 * 10^{-3}$

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	$0.005 \mid 1.6 * 10^{-4}$	$0.017 \mid 4.7 * 10^{-4}$	$0.039 \mid 8.6 * 10^{-4}$
DSM	$0.017 \mid 5.0 * 10^{-4}$	$0.016 \mid 1.1 * 10^{-3}$	$0.038 \mid 2.5 * 10^{-3}$

Table 4. Brier scores (first number in each entry) and average variances of $S(t)$ (second number in each entry) for RDSM and DSM at different event horizons for composite (upper panel) and mortality (lower panel) outcome.

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	0.788	0.760	0.722
DSM	**0.803**	**0.776**	**0.727**
GBM	0.798	0.768	0.725
RSF	**0.803**	0.766	0.714
Cox	0.788	0.757	0.710

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	**0.790**	**0.782**	0.711
DSM	0.763	0.780	**0.715**
GBM	0.759	0.776	0.713
RSF	0.740	0.759	0.709
Cox	0.739	0.767	0.708

Table 5. Age Subgroup (age 65-75) analysis for the composite (left) and PC-mortality (right) outcome.

outcome included a specific value of the PSA test (PSA > 50 ng/ ml). Nevertheless, PSA-related features had a significant influence on the model performances for the morality outcome as well.

Regional Analysis

One of the most important criteria to assess a model's generalizability is independent external validation. Unfortunately, we had no such data at our disposal. Instead, we devised a proxy method to approximate independent external validation. According to the definition of the United States Census Bureau, we first divided all VA facilities into 4 main regions, i.e., Northeast, Midwest, South, and West. Each region accounted for 13%, 22%, 44%, and 19% of our cohorts[1], respectively. Then we trained our model using data from 3 regions and reserved the patients from the left-out region as the validation group. Table 8 shows the validation results using patients from the Northeast region. Supplemental materials contain the results for the other three regions. Compared with Table 3, where patients from different regions were mixed, we found that all models generalized well geographically.

One noticeable exception occurred in the Midwest region (Table 10), where all models performed significantly worse than other regions for the 2-yr composite outcome predictions. We found the average PSA values were slightly higher (9.3 vs. 8.9 ng/ml) than the rest of the regions. While other important features (Gleason score, Clinical T) were quite similar. The PSA difference is unlikely to be solely responsible for the anomalous results. We will leave the detailed investigation to the future work.

Discussion

In this study, we developed and tested different survival models to predict localized prostate cancer prognosis in 2, 5, and 10-yr time horizons, using routinely available EHR data from the largest integrated healthcare system in the US. In addition to the conventional PC-mortality outcome, we also considered a composite outcome, which encompasses the PSA values, metastasis, and PC-mortality. Overall, in terms of the time-dependent concordance

[1] We treated patients from other US territories as a single group and put them into the training data. They represented 2% of our total population.

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	0.839	0.782	0.739
DSM	0.844	0.782	0.738
GBM	0.847	0.790	**0.743**
RSF	**0.851**	**0.792**	0.740
Cox	0.839	0.769	0.730

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	0.867	0.841	0.781
DSM	**0.875**	**0.851**	0.797
GBM	0.845	0.834	0.790
RSF	0.843	0.835	0.797
Cox	0.874	0.846	**0.808**

Table 6. Race Subgroup (black) analysis for the composite (left) and PC-mortality (right) outcome.

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	0.850	0.797	0.755
RDSM w/o interval	0.840	0.793	0.755
RDSM w/o race	0.848	0.787	0.750
RDSM w/o age	0.845	0.789	0.744
RDSM w/o stage	0.837	0.764	0.718
RDSM w/o psa	0.743	0.730	0.716
DSM	0.853	0.796	0.758
DSM w/o race	0.850	0.794	0.755
DSM w/o age	0.842	0.785	0.739
DSM w/o stage	0.831	0.762	0.718
DSM w/o psa	0.736	0.731	0.717

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	0.851	0.836	0.805
RDSM w/o interval	0.827	0.839	0.796
RDSM w/o race	0.831	0.834	0.793
RDSM w/o age	0.816	0.826	0.788
RDSM w/o stage	0.795	0.789	0.760
RDSM w/o psa	0.790	0.797	0.762
DSM	0.821	0.824	0.789
DSM w/o race	0.812	0.820	0.785
DSM w/o age	0.798	0.815	0.776
DSM w/o stage	0.793	0.777	0.747
DSM w/o psa	0.780	0.790	0.754

Table 7. Ablation study results for the composite (left) and PC-mortality (right) outcome. For reference, we also include results with all data on the top row of the table.

index C_{td}, two deep learning models RDSM and DSM, outperformed traditional machine learning models and the Cox model with a moderate margin (Table 3). We also noticed that DSM and RDSM were much faster than RSF and GBM in training and inference. In the short term scenario (2-yr), all models yielded higher C_{td} for the composite outcome than the PC-mortality outcome. We attribute this result to the low PC-mortality rate within the first 2-yr of diagnosis, resulting in a more skewed outcome distribution. For the longer period (5 and 10-yr), all models experienced considerable performance drop for the composite outcome. And the C_{td} became lower than the corresponding PC-mortality prediction. The results suggest that our composite outcome, while being more clinically relevant for the short-term prognosis management, poses greater challenges for the long-term prediction.

For the PC-mortality outcome, our best model RDSM achieved the C-index 0.807 at the 10-yr time horizon, which is comparable to other contemporary studies[5,7,8,10] using multi-variable approaches with similar input variables. Although we should address that our results were not validated with external independent data. To compensate for this limitation, we conducted a regional analysis by dividing the population into geographically distinct areas and using the left-out region as the test set. With one exception for the Midwest region, all models demonstrated reasonable generalizability. We will leave a detailed investigation for future work.

We further performed analysis on subgroups stratified by age and race. We found that the performance variances among different age groups were higher than in race groups, suggesting that age is a more important predictor than race. We conducted a thorough ablation study and identified that prostate cancer stage information and PSA-related features were the most important features for both composite and PC-mortality outcomes. We noticed that the impact of prostate cancer stages (Gleason score and Clinical T) increased over time, while PSA-related features mattered the short-term (2-yr) predictions the most. The ablation study also justifies the inclusion of PSA values in our definition of the composite outcome. Another interesting observation from the ablation study was that the

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	**0.862**	**0.824**	**0.765**
DSM	0.856	0.822	**0.765**
GBM	0.856	0.822	0.761
RSF	0.853	0.818	0.756
Cox	0.858	0.819	0.753

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	**0.881**	**0.863**	**0.807**
DSM	0.870	0.844	0.799
GBM	0.847	0.843	0.792
RSF	0.850	0.841	0.789
Cox	0.855	0.844	0.790

Table 8. Validation results for the composite (left) and PC-mortality (right) outcome using the Northeast cohort.

longitudinal model RDSM didn't benefit much from the time interval information. We conjecture it is because the neural network architecture of RDSM, Recurrent Neural Network (RNN), is a discrete sequential model, which can be inefficient to learn continuous and irregularly spaced temporal signals without further modification[11].

One difference in our study with some previous work[5,6] is that we didn't include treatment methods, e.g., hormone therapy and radiotherapy, as input variables in our models. Our focus is to provide an initial risk estimation before giving any treatment.

Traditional statistical survival analysis can be roughly divided into three categories: non-parametric, semi-parametric, and parametric[12]. The semi-parametric Cox model, and its variants, have been widely adopted in the clinical survival analysis. The proportional hazards assumption of the Cox model can oftentimes be violated, especially when the event horizon is long. On the other hand, the parametric method, by assuming the survival times follow a particular distribution, is efficient and easy to interpret. However, its performance will suffer if the underlying distribution deviates significantly from the prior distribution. To overcome the limitation, we employed an ensemble approach by combining a large number of parametric regression models, where their combination weights and distribution parameters are learned via deep neural networks. The combination of neural network models and the analytical parametric approach enhances the model performance while preserving interpretability. As a natural extension of ensemble learning, our deep learning models can provide reliable uncertainty quantification. We showed that weighted variance of the predicted survival probability $S(t)$ consistently correlated with the Brier score (Table 4). The reliable UQ will help clinicians make better-informed decisions.

The major strength of our study is that we made use of a large cohort of localized prostate cancer patients from the VA national medical system. And we tested the various survival models as the benchmark. Our deep learning models demonstrated sufficient discriminating power plus the capability of uncertainty quantification. The main limitation of our study was the lack of independent external validation. Thus we couldn't uncover the potential difference between veterans and the general population which might impact the performance of real-world applications.

To conclude, our novel deep learning approach, equipped with uncertainty quantification, can provide accurate individualized risk estimation for localized prostate cancer patients. Our study may further motivate the implementation of an AI-assisted clinical decision system for prostate cancer prognosis management.

Methods

Population Selection and Data Processing

We used the AJCC TNM staging system[13] to select localized prostate cancer patients. TNM denotes the extent of the primary tumor (T), whether cancer has spread to nearby lymph nodes (N) and distant parts of the body (M). For prostate cancer, the TNM staging system also considers the PSA level at the time of diagnosis and the Gleason score, which measures how likely the cancer is to grow and spread. By definition, for all stage I and II patients, the primary tumor is localized at the time of diagnosis.

Given the importance of the PSA test for prostate cancer, we collected all the PSA test results, up to 10 years before diagnosis for each patient. Other clinical input features include the Gleason score (range from 6-10), and the clinical prostate tumor stage, i.e., T-stage in the TNM staging system. Finally, the patient's age and race information were also included. To sum up, the input data for our longitudinal model RDSM consisted of time-ordered PSA tests, age at each PSA test, and the time distance between each test and the time of diagnosis, plus the Gleason score, T-stage and the race. Whereas for time-independent models, we replaced the PSA tests with their summary statistics, which included the number of PSA tests, maximum, minimum, average, last, and penultimate PSA values before diagnosis, and a binary indicator showing whether the last PSA test elevated compared with the penultimate test.

Definition of Patient Outcomes and Evaluation Metrics

The relatively long survival time and low mortality rate of localized prostate cancer pose a great challenge in risk estimation. To get a more accurate disease prognosis evaluation over a shorter and practical time scale, we define a composite outcome:

- Prostate-Specific Antigen (PSA) > 50 ng/ml
- Metastatic diseases
- Prostate cancer mortality

as our event of interest. The event time is the earliest date of any of these three events. The censoring time is 1 year after the last PSA test. In the cases that patients died of other causes before censoring, the censoring time is the time of death instead.

To get a better insight into the model performances over time, we calculate the time-dependent concordance-index $C_{\text{td}}(t)$[14]:

$$C_{\text{td}}(t) = \mathbb{P}\left(\hat{F}\left(t \mid \mathbf{x}_i\right) > \hat{F}\left(t \mid \mathbf{x}_j\right) \mid \delta_i = 1, T_i < T_j, T_i \leq t\right) \tag{1}$$

at some truncation time t. Here $\hat{F}\left(t \mid \mathbf{x}_i\right)$ is the cumulative distribution function (CDF) at time t, given input feature $\mathbf{X}$. To account for the high censoring ratio, we adjust $C_{\text{td}}(t)$ with the inverse probability of censoring weights[15]. Additionally, we test our models against the more conventional outcome, namely, prostate cancer mortality. In this study, we set the truncation time to be 2, 5, and 10 yrs after diagnosis.

Depending on whether the input $\mathbf{X}$ is time-dependent, we employed two different deep learning models, Deep Survival Machine (DSM)[16] and Recurrent Deep Survival Machine (RDSM)[17]. As a benchmark, we also considered two popular machine learning models, Random Survival Forest (RSF)[18] and Gradient Boosting Machine (GBM) [19], and the classical Cox model[20–23]. And we implemented all three benchmark models using the scikit-survival package[24].

$C_{\text{td}}(t)$[14] is a good metric in gauging the model performance. However, by definition, it involves pairwise comparisons between different individuals. For the sake of UQ, it is more appropriate to evaluate the prediction accuracy using the Brier score[25], which is defined as:

$$\text{BS}(t) = \frac{1}{n}\sum_{i=1}^{n} I\left(y_i \leq t \wedge \delta_i = 1\right) \frac{\left(0 - S\left(t \mid \mathbf{x}_i\right)\right)^2}{\hat{G}\left(y_i\right)} + I\left(y_i > t\right) \frac{\left(1 - S\left(t \mid \mathbf{x}_i\right)\right)^2}{\hat{G}(t)} \tag{2}$$

where $S\left(t \mid \mathbf{x}_i\right)$ is the predicted survival function with input x, $I(\cdot)$ is the indicator function, and $1/\hat{G}$ is an inverse probability of censoring weight, estimated by the Kaplan-Meier estimator. By the definition in Eq. (2), Brier score measures the distance between the predicted survival probability with the true survival status, weighted by the censoring weights.

Model Overview

Parametric models assume that the survival times or the logarithm of the survival times of the population follow a particular distribution. For example, the Weibull distribution, which is one of the most used distributions for the survival analysis, is characterized by the shape parameter k and the scale parameter λ. The corresponding hazard function takes the form $h(t) = \lambda k t^{k-1}$, which means that the risk can change over time depending on the value of k.

Compared with the semi-parametric Cox model, the parametric model is free of the proportional hazards assumption, which may not be realistic in our case, as the time to event is long and risk can vary over time.

However, for the large-scale EHR data collected over a long period and across different locations, a single distribution function may be insufficient to characterize the whole population without introducing biases toward some specific sub-groups. Besides, we argue that a single distribution function contains too few parameters to utilize the rich information embedded in the large heterogeneous dataset. Thus, to increase the model capacity and reduce the potential bias, we adapt a deep learning-augmented ensemble approach, which was first introduced in Ref.[16,17]. The key idea is to model the conditional survival function $\mathbb{S}(t \mid X) \triangleq \mathbb{P}(T > t \mid X)$ as an *ensemble* of parametric distributions, while parameters of each distribution function and mixing weights are all learned from a neural network. This ensemble approach can help lower the variance and increase the out-of-sample performance. And the expressive power of deep learning models allows us to make more efficient use of patients' data.

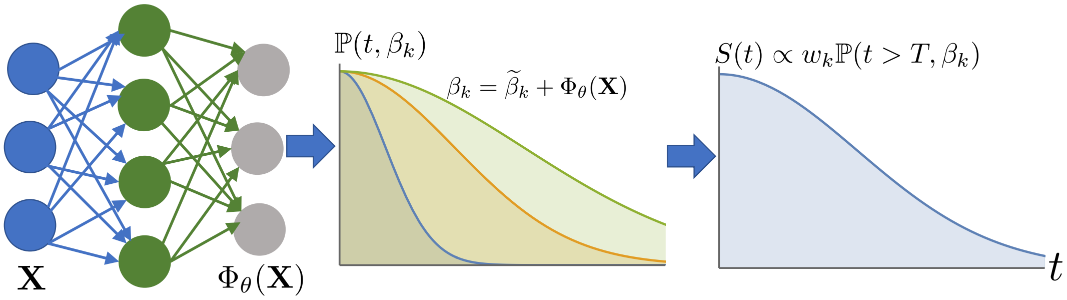
$\mathbf{X}$
$\Phi_\theta(\mathbf{X})$
$\mathbb{P}(t, \beta_k)$
$\beta_k = \tilde{\beta}_k + \Phi_\theta(\mathbf{X})$
$S(t) \propto w_k \mathbb{P}(t > T, \beta_k)$
t

Figure 3. Model overview. The neural network is responsible for learning the feature representation $\Phi_\theta(\mathbf{X})$ given the input $\mathbf{X}$. The parameters of all k distributions β_k, and their mixing weights w_k, are learned jointly during the training.

To begin with, we choose a distribution function $\mathbb{P}(t, \beta)$, preferably the Log-Normal or Weibull distribution, because their closed form of CDF can simplify the gradient descent optimization. Here we use β to denote all the relevant parameters of the distribution function. Then we initialize k such distributions with some random prior parameters $\tilde{\beta}_k$. The input features x_i of patient i, are passed through the neural network Φ_θ to find the representation $\Phi_\theta(x_i)$.

More precisely, we have

$$\beta_k = \tilde{\beta}_k + \Phi_\theta(x_i) \tag{3}$$

The mixture weights w_k are also learned jointly via the optimization procedure outlined below. Depending on whether the input data is time-dependent or not, Φ_θ can be either a Recurrent Neural Network (RNN) or a Multi-Layer Perceptron (MLP). And we call them RDSM (Recurrent Deep Survival Machine) and DSM (Deep Survival Machine), respectively[16,17]. In practice, we choose either LSTM[26] or GRU[27] as the concrete realization of the RNN module in the RDSM model. The training proceeds by calculating the maximum likelihood estimator, which amounts to minimizing the following loss function:

$$\mathscr{L}_{\text{combined}} = \mathbf{ELBO}_U(\Theta) + \alpha \cdot \mathbf{ELBO}_C(\Theta) \tag{4}$$

where the first term denotes the uncensored loss,

$$\mathbf{ELBO}_U = \sum_{i=1}^{|\mathscr{D}|} \left(\underset{Z \sim (\cdot|\mathbf{x}_i, w)}{\mathbb{E}} [\ln \mathbb{P}(T = t_i \mid Z, \beta_k,)] \right), \tag{5}$$

and the second term is for the censored loss,

$$\mathbf{ELBO}_C = \sum_{i=1}^{|\mathscr{D}|} \left(\underset{Z \sim (\cdot|\mathbf{x}_i, w)}{\mathbb{E}} [\ln \mathbb{P}(T > t_i \mid Z, \beta_k)] \right). \tag{6}$$

Here $\alpha \in [0,1]$ is a discount factor and we treat it as a hyperparameter, and w denotes the mixture weight. To mitigate the long-tail bias, we add L_2 regularization for β_k during the training. The final survival probability $S(t|X)$ is the weighted average over k distributions:

$$S(t \mid X) = \sum_{w} w_k \mathbb{P}_k(T > t \mid X, \beta_k) \tag{7}$$

Data Availability

The data used in this study cannot be made available due to restrictions relating to the use of EHR data.

Code Availability

The code used in this study is available from the authors upon request.

Acknowledgements

This study was funded by the Million Veteran Program, Office of Research and Development, Veterans Health Administration (MVP000 and MVP017). This publication does not represent the views of the Department of Veteran Affairs or the United States Government.

Supplemental Material

The supplemental material contains results of all regional and subgroup analysis. It is worth mentioning that we use the same parameters as the general case to ensure a fair comparison. We expect that if we perform hyper-parameter tuning in each case, we could see moderate performance increases for all models. However, the training of two traditional machine learning models, GBM and RSF, takes significantly larger computational resources than the rest. For example, training a RSF model on the whole training set (101,048 patients) takes more than 24h to finish, while two deep learning models (RDSM and DSM) only require a few minutes.

Regional Study Results

Table 9, 10 and 11 show the validation results using patients from South, Midwest and West regions.

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	0.824	0.788	0.742
DSM	0.836	0.795	0.750
GBM	0.838	**0.796**	**0.752**
RSF	**0.840**	**0.796**	0.750
Cox	0.833	0.791	0.745

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	0.813	0.805	0.772
DSM	0.823	**0.815**	**0.779**
GBM	0.827	0.812	0.777
RSF	**0.828**	0.812	0.773
Cox	0.817	0.809	0.774

Table 9. Validation results for the composite (left) and PC-mortality (right) outcome using the South cohort.

Subgroup Analysis Results

The stratification of different age and race subgroups can be found in Fig. 2 of the main text. Table 12, 13, 14 show the result for three different age groups. Table 15 and 16 show the results for the white and the other race groups.

References

1. Cancer stat facts: Prostate cancer. https://seer.cancer.gov/statfacts/html/prost.html.
2. Wilt, T. J. *et al.* Radical prostatectomy versus observation for localized prostate cancer. *N Engl J Med* **367**, 203–213 (2012).

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	0.761	**0.764**	0.737
DSM	0.756	0.763	**0.740**
GBM	0.760	0.763	0.736
RSF	**0.764**	0.762	0.723
Cox	**0.764**	0.754	0.724

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	**0.848**	**0.838**	**0.798**
DSM	0.830	0.826	0.792
GBM	0.817	0.824	0.789
RSF	0.813	0.817	0.786
Cox	0.823	0.817	0.781

Table 10. Validation results for the composite (left) and PC-mortality (right) outcome using the Midwest cohort.

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	0.875	0.803	0.747
DSM	**0.882**	0.809	0.750
GBM	0.878	**0.810**	**0.752**
RSF	0.878	0.805	0.743
Cox	0.863	0.797	0.737

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	**0.865**	**0.827**	**0.784**
DSM	0.852	**0.827**	**0.784**
GBM	0.857	0.821	0.777
RSF	0.855	0.817	0.767
Cox	0.820	0.810	0.773

Table 11. Validation results for the composite (left) and PC-mortality (right) outcome using the West cohort.

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	0.778	0.767	0.714
DSM	0.821	0.813	0.716
GBM	0.824	0.813	0.715
RSF	**0.839**	**0.830**	**0.728**
Cox	0.812	0.828	0.719

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	N/A	**0.886**	0.827
DSM	N/A	0.884	**0.835**
GBM	N/A	0.788	0.773
RSF	N/A	0.794	0.794
Cox	N/A	0.863	0.816

Table 12. Age Subgroup (age < 55) analysis for the composite (left) and PC-mortality (right) outcome. The 2-yr C-index is not applicable for the PC-mortality as all subjects in this age group are censored.

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	0.829	0.801	0.733
DSM	0.833	0.799	0.728
GBM	0.829	**0.806**	**0.740**
RSF	**0.838**	0.804	0.737
Cox	0.830	0.796	0.729

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	**0.852**	**0.754**	0.734
DSM	0.804	0.745	0.756
GBM	0.786	0.733	**0.760**
RSF	0.789	0.745	0.752
Cox	0.779	0.738	0.752

Table 13. Age Subgroup (age 55-65) analysis for the composite (left) and PC-mortality (right) outcome.

3. Hamdy, F. C. *et al.* 10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. *N Engl J Med* **375**, 1415–1424 (2016).

4. Stephenson, A. J. *et al.* Prostate cancer–specific mortality after radical prostatectomy for patients treated in the prostate-specific antigen era. *J. Clin. Oncol.* **27**, 4300 (2009).

5. Thurtle, D. R. *et al.* Individual prognosis at diagnosis in nonmetastatic prostate cancer: Development and external validation of the predict prostate multivariable model. *PLoS medicine* **16**, e1002758 (2019).

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	0.814	0.770	0.714
DSM	**0.829**	0.774	0.720
GBM	0.823	**0.775**	**0.721**
RSF	0.823	0.772	0.716
Cox	0.822	0.764	0.715

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	**0.801**	0.755	0.753
DSM	0.796	0.765	0.750
GBM	0.787	**0.774**	**0.762**
RSF	0.800	0.770	0.752
Cox	0.787	0.765	0.753

Table 14. Age Subgroup (age > 75) analysis for the composite (left) and PC-mortality (right) outcome.

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	**0.833**	0.796	**0.760**
DSM	0.814	0.803	0.749
GBM	0.815	**0.806**	0.751
RSF	0.820	**0.806**	0.751
Cox	0.799	0.788	0.733

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	**0.814**	0.830	**0.781**
DSM	0.807	0.828	0.771
GBM	0.802	0.831	0.776
RSF	0.820	**0.834**	0.773
Cox	0.789	0.831	0.771

Table 15. Race Subgroup (white) analysis for the composite (left) and PC-mortality (right) outcome.

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	0.773	0.800	0.745
DSM	0.816	**0.824**	**0.751**
GBM	0.793	0.801	0.741
RSF	**0.826**	0.813	0.750
Cox	0.807	0.799	0.749

Event Horizon	2-yr	5-yr	10-yr
Model			
RDSM	0.856	0.825	0.700
DSM	0.878	**0.887**	**0.735**
GBM	0.863	0.834	0.709
RSF	**0.897**	0.803	0.672
Cox	0.883	0.876	0.734

Table 16. Race Subgroup (other) analysis for the composite (left) and PC-mortality (right) outcome.

6. Bibault, J.-E. *et al.* Development and validation of an interpretable artificial intelligence model to predict 10-year prostate cancer mortality. *Cancers* **13**, 3064 (2021).

7. Lee, C. *et al.* 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. *The Lancet Digit. Heal.* **3**, e158–e165 (2021).

8. Zelic, R. *et al.* Predicting prostate cancer death with different pretreatment risk stratification tools: a head-to-head comparison in a nationwide cohort study. *Eur. urology* **77**, 180–188 (2020).

9. Park, J. H. *et al.* Machine learning prediction of incidence of alzheimer's disease using large-scale administrative health data. *NPJ digital medicine* **3**, 1–7 (2020).

10. Dess, R. T. *et al.* Development and validation of a clinical prognostic stage group system for nonmetastatic prostate cancer using disease-specific mortality results from the international staging collaboration for cancer of the prostate. *JAMA oncology* **6**, 1912–1920 (2020).

11. Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Sci. reports* **8**, 1–12 (2018).

12. Wang, P., Li, Y. & Reddy, C. K. Machine learning for survival analysis: A survey. *ACM Comput. Surv. (CSUR)* **51**, 1–36 (2019).

13. Amin, M. B. *et al.* The eighth edition ajcc cancer staging manual: continuing to build a bridge from a population-based to a more “personalized” approach to cancer staging. *CA: a cancer journal for clinicians* **67**, 93–99 (2017).

14. Antolini, L., Boracchi, P. & Biganzoli, E. A time-dependent discrimination index for survival data. *Stat. medicine* **24**, 3927–3944 (2005).

15. Uno, H., Cai, T., Pencina, M. J., D’Agostino, R. B. & Wei, L.-J. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. medicine* **30**, 1105–1117 (2011).

16. Nagpal, C., Li, X. R. & Dubrawski, A. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE J. Biomed. Heal. Informatics* (2021).

17. Nagpal, C., Jeanselme, V. & Dubrawski, A. Deep parametric time-to-event regression with time-varying covariates. In *Survival Prediction-Algorithms, Challenges and Applications*, 184–193 (PMLR, 2021).

18. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *The annals applied statistics* **2**, 841–860 (2008).

19. Ridgeway, G. The state of boosting. *Comput. science statistics* 172–181 (1999).

20. Cox, D. R. Regression models and life-tables. *J. Royal Stat. Soc. Ser. B (Methodological)* **34**, 187–202 (1972).

21. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. royal statistical society: series B (statistical methodology)* **67**, 301–320 (2005).

22. Benner, A., Zucknick, M., Hielscher, T., Ittrich, C. & Mansmann, U. High-dimensional cox models: the choice of penalty as part of the model building process. *Biom. J.* **52**, 50–69 (2010).

23. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for cox’s proportional hazards model via coordinate descent. *J. statistical software* **39**, 1 (2011).

24. Pölsterl, S. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *J. Mach. Learn. Res.* **21**, 1–6 (2020).

25. Graf, E., Schmoor, C., Sauerbrei, W. & Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Stat. medicine* **18**, 2529–2545 (1999).

26. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735–1780 (1997).

27. Cho, K. *et al.* Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).