Artificial Intelligence-based Personalized and Risk-adapted Surveillance Management for Urologic Cancer: A SEER-based Study

Okyaz Eminaga (1,2), Bernhard Breil (5), Axel Semjonow (3), Martin Boegemann (3), Lei Xing (4), Ilker Tinay (6), and Joseph C. Liao (1,2)

(1) Department of Urology, Stanford University School of Medicine
(2) Center for Artificial Intelligence in Medicine & Imaging, Stanford University School of Medicine
(3) Prostate Center, Dept. of Urology, University Hospital Muenster, Germany
(4) Department of Radiation Oncology, Stanford University School of Medicine
(5) Faculty of Health Care, Hochschule Niederrhein, University of Applied Sciences, Krefeld, Germany.
(6) Department of Urology, Marmara University School of Medicine, Istanbul, Turkey

Corresponding author:
Okyaz Eminaga, M.D./Ph.D.
Department of Urology, Center for Artificial Intelligence in Medicine & Imaging (AIMI), Stanford University School of Medicine
300 Pasteur Drive, S-287
Stanford CA 94305-5118
Tel: 650-725-5544
Fax: 650-723-0765
Email: okyaz.eminaga@stanford.edu
Abstract:

Identification of prognosis changing events is essential during cancer surveillance. Current clinical practice is limited by “one-size-fits-all” time management or subjective assessment that is not necessarily appropriate for individual patients’ risk profiles. A risk-adapted follow-up is more efficient and optimal from the clinical perspective. Herein we introduce an artificial intelligence-based data-driven solution for risk-adapted follow-up schedule that devises the number of follow-up visits per year and assesses the cancer-specific risk profile over a long interval depending on the age at diagnosis. We utilized Surveillance, Epidemiology, and End Results (SEER), a cancer registry database and a resource for cancer-specific survival estimation in the United States that includes more than 2 million patients diagnosed with urologic cancers. We tested different machine learning algorithms on their definition for the feature importance and selected clinically relevant parameters regularly used in clinical routine to develop a survival modeling. As a result, the underlying recurrent neural network algorithm for follow-up modeling was fitted on the unseen test set with an overall concordance index score of 0.80. A controlled access to the online tool is available for physicians.
Introduction

Devising an individualized surveillance strategy for patients following oncologic interventions is still one of the most challenging tasks for clinicians caring for cancer patients. The treatment and surveillance strategy are typically based on the best possible risk profiles and patients associated with comorbidities \(^1\). After curative treatment or during active surveillance, patients are followed at predefined intervals to identify treatment-associated side effects or prognosis changing events that require the clinical reassessment, initiating a new treatment strategy or modifying the ongoing treatment \(^2\). The follow-up visits generally start three months after treatment and are conducted at 3 to 6-month intervals initially and possibly extended to an annual basis depending on cancer risk and time interval from the initial treatment \(^3\). The frequency of the follow-up visits is generally based on expert opinions, given that some patients will have a failed curative treatment in the first years after diagnosis \(^4\). The current clinical practice is limited by rigid time management or the subjective assessment that is not necessarily appropriate for the patients’ risk profile.

We hypothesize that risk-adapted individualized follow-up planning for cancer diseases is feasible. Towards that goal, we introduced artificial intelligence (AI)-based data-driven solution that recommends the number of follow-up visits per year and assesses the cancer-specific risk profile over a long period. Further, it suggests when we shall reconsider the follow-up plan. For that purpose, we utilized a cancer registry database, including almost 2 million patients who were diagnosed with cancer of urogenital organs (i.e., prostate, testis, kidney, urinary bladder, ureter, renal pelvis and penis) from the Surveillance, Epidemiology, and End Results (SEER) Program that is considered as resources for the cancer-specific survival estimation in the United States \(^5\). We further
included a web-based explainable AI solution interface to recommend a risk-adapted follow-up strategy using an advanced machine learning algorithm and the puzzle concept for urologic cancer cases. We utilized the puzzle concept that assumes that all patients from a national dataset are pieces of a puzzling problem to reconstruct the population's survival history. Our concept relies on the recurrent artificial neural network that supports solving the N-puzzle problem and is frequently applied to natural language processing (e.g., automated generation of texts)

Material and Methods

Dataset
We utilized the SEER database covering the years 1975-2017 (Version 18) and included 2,006,052 patients diagnosed with one of the urologic cancers (i.e., prostate, testis, kidney, urinary bladder, ureter, renal pelvis, penis, other genital organs). The affected organ by cancer type was determined according to the documentation guideline of the SEER program (‘Site recode B ICD-O-3/WHO 2008’). Here, we collected 505 features from the database covering demographic information, clinical and pathological information. Information on treatments was not included, since only surgery data are available. Further, two additional features (cancer-specific death status and the follow-up duration given in years) included the cancer-specific survival follow-up data. Supplement file 1 lists the variables considered in the current study.

We randomly divided the database into the development set (90%) and the out-held test set (10%) while maintaining the same proportion for the cancer-specific death status between these datasets for model development and evaluation. The development set was then split into the training (90%) and validation sets (10%). The validation set is used to evaluate the optimization procedure of the model weights. We applied the feature-wise
normalization by ranking the units of a feature according to their appearance order in the dataset and dividing by the maximum rank to achieve a value between 0 and 1. When the information is missing for the feature, a value of -1 was given. In contrast, the follow-up duration was fed as a non-negative continuous value without normalization into the model, and the cancer-specific death status was binarized (0: no cancer-specific death; 1: cancer-specific death).

**Reliability of Machine Learning for the feature selection**

A total of 506 features including the follow-up duration were considered as input data. We evaluated the feature importance to determine whether machine learning can provide a list of clinically relevant features. For all features, we evaluated the Gini importance for the random forest and XGBoost algorithms and the coefficients for the Linear Support Vector Machine (LSVM). Besides, we estimated the coefficients from the hidden layer and the Shapley values for the simple recurrent neural network (sRNN). Shapley values evaluate the contribution of each feature to the prediction. To estimate the coefficients from the hidden layer for the simple recurrent neural network, we applied the feature-wise summation of the coefficients (506 different coefficients were available per feature). These models were trained on the training set and optimized on the validation set, and the feature importance was determined on the test set. Here, the input data were applied to determine the binarized cancer-specific death status.

Finally, we evaluated whether all top 20 features for each algorithm were clinically relevant from the urological perspective or included administrative, social, and demographic parameters or outdated features no longer considered by the SEER program. The clinical relevance was determined based on whether this feature covered the tumor staging, race, age at diagnosis, the number of malignancies, treatment-related information, or a cancer-specific risk factor (e.g., biopsy information, biomarkers, number
of positive lymph nodes). Further, the list shall not contain redundant or outdated features (e.g., year of diagnosis, different versions of parameters for tumor stage). A top feature list was considered invalid when it contained any clinically not relevant features.

**Feature Selection for Model Development**

We aimed to select features for survival modeling that have the relevant clinical information covering tumor stage and biomarkers in addition to age at diagnosis and race due to their well-established clinical importance in urologic cancers 1-5.

The approach for feature selection was considered depending on the results from the previous section. If all top feature lists were invalid, we manually selected the clinically relevant features which are currently considered by the SEER program. While selecting the features, we intended to reduce the version dependency of the tumor stage in our model by considering features that describe the tumor extension and tumor dissemination in more detail. Further, the SEER tumor stage was considered given that the SEER program used this feature to estimate the cancer-specific survival rate according to the tumor dissemination status.

These features were selected in full agreement of 5 urologists (OE, AS, MB, JL, IT) who have a profound clinical and research experience in urologic cancers for more than 10 years. The feature number would depend on the result of the feature selection procedure.

**Model Development for the risk-profile reconstruction**

Since the SEER database can cover all variations of the disease conditions and characteristics6, we defined a puzzle concept to develop our cancer-specific survival model. For that purpose, we preferred the recurrent neural network (RNN) to reconstruct the risk profile based on information sequence since RNN is effective in learning the prediction of the outcome based on the information sequence 6. For model development, we assumed that each patient provided an information sequence for the given t time and
therefore contributed to the reconstruction of the cohort’s risk profile history. The model would be able to estimate the cancer-specific survival probabilities for a period $T$ where $t$ is an element of $T: = [1,44]$. While developing and validating our model, we followed the recommendation made by PROGRESS.

**Hyperparameter Configuration**
A grid search was applied to identify the optimal hyperparameter configuration for model architecture [i.e., simple recurrent neural network (sRNN), gated recurrent units (GRU) and Long Short-Term Memory unit (LSTM)] and model optimization algorithms with the default configuration [i.e. Adaptive Moment Estimation (Adam) and Root Mean Square Propagation (RMSprop)]. The log-likelihood function was measured to evaluate the model optimization with a predefined decay learning rate (start learning rate: 0.001 and decay rate: 0.8 after every fifth epoch). Further, we evaluated the inclusion of the dropout in the model on the log-likelihood. The linear regression function was applied for the final activation function, and the mean square error was used as a loss function.

**The risk velocity-based follow-up assessment**
After developing the risk-profile reconstruction model, we estimated the risk velocity as risk change over a time unit because it reflects the time-dependent alteration in the risk profile and prognosis of the case. High velocity means a fast change of the prognosis, and a low velocity means that the risk profile is changing slowly over time. A velocity stillstand means that the risk profile is stable over time.

To develop the risk velocity-based follow-up assessment for a case, we defined the following algorithms:
1. Since we aimed to develop a risk-adapted follow-up strategy for a single case, the risk prognostication is considered as a follow-up score \((s)\). The collection of the follow-up scores \((S)\) are defined as \(S = \{s \mid s \in \mathbb{R} \land 0 \leq s \leq 1\}\).

2. We defined a set of 120 time points (data points) \((P)\) with an equal time distance in a year \((k)\) in the period \((f)\) from the age at diagnosis \((a)\) to the median U.S. life expectancy\(^{a} (b)\) (78 years) to generate a smooth follow-up curve.
   
   i. \(F = \{f \mid f \in \mathbb{N}^{+} \land f = a - b \mid a, b \in \mathbb{N}^{+} \land a < b \}\).
   
   ii. \(k = \frac{f}{120}\) where \(f \in F\)
   
   iii. \(P = \{p \in S \mid |P| = 120\}\) and \(A = \{kn\}_{n=0}^{119}\) where \(k\) is a constant from ii and \(A\) is a set of time positions for the data points \(P\), thereby \(P \Leftrightarrow A\).

3. Then, we assigned data points \((P)\) to a set of time intervals \((l)\), thereby resulting in sequences of interval-associated data points \((D)\). The interval length \((l)\) is dynamic and corresponds to the half number of data points of a year (e.g., for a 12 years long follow-up period, the interval length would be 5 data points). The number of sequences \((n)\) corresponds to the time intervals.

   iv. \(l = \frac{120}{2f}\) where \(f \in F\)

   v. \(n = \frac{120}{l}\)

   vi. \(l = \{(li, l(i + 1))\}_{i=0}^{n-1}\) where \(l\) is a constant and calculated using iv

4. The score velocity \((v)\) is the difference between the first \((p_{1})\) and the last \((p_{m})\) scores \([p_{1}, p_{m}]\) of a sequence from \(D\).

   vii. \(v = |p_{1} - p_{m}|\)
5. Since the score velocity reflects the dynamic of the risk profile, we decided to derive the number of follow-up visits from the score velocities of a year using the formula:

\[ \text{viii. The visit number for a year} = \begin{cases} \min \left( \sum_{l=0}^{n} |v_l| \times 1000 + 1, 4 \right), & \sum_{l=0}^{n} |v_l| > 0 \\ 0, & \sum_{l=0}^{n} |v_l| \leq 0 \end{cases} \]

We limited the maximum visit number per year to 4 which is generally the maximum number of follow-up visits per year for urologic cancers. Only score velocities above 0.01% were considered for the calculation. For each year, we summed the absolute values of score velocities as given above. The number of score velocities in a year was estimated as following:

\[ \text{ix. } n = 2l. \]

6. The definition of the period for a close follow-up is based on the grade of velocities; time intervals with score velocities exceeding the threshold of 0.5% were marked for a close follow-up. The reason for defining 0.5% is that the overall lowest 5-year cancer-specific death probability was 0.5%.

7. The follow-up plan needs to be adapted over time. Therefore, we added a suggestion rule based on the score velocity to define when the follow-up plan to reconsider. Since a score velocity of 0 means that the risk profile remained unchanged or stable for a given time interval, it is intuitively to assess the follow-up plan in this condition.

**Evaluation Metrics**

The discriminatory accuracy was measured by Harrell’s Concordance Index. The fitness of the model was assessed by comparing Kaplan–Meier Curves between the prognosticated and observed probabilities. The clinical utility of survival modeling was
assessed by the capability to define discriminative risk groups. Here, we assessed the observed cancer-specific survival probabilities at the 10\textsuperscript{th} year from diagnosis after stratifying the test set by quantile thresholds (10\%, 50\%, 90\%) determined from the training set for the prognosticated cancer-specific death probability. The log-rank test was applied to determine the discrimination significance between the two risk groups. The two-sided significance is reached when $P<0.05$. Finally, we demonstrated our solution on four patients seen at Prostate Cancer Center of University Hospital Muenster in Germany who had longitudinal information between 3-14 years and are not included in the SEER database.

**Machine Learning Framework and Statistical Tools**

The model development and evaluation were based on Python 3.6 (Python Software Foundation, Wilmington, DE) and applied the Shapley, scikit-learn, TensorFlow, and Keras libraries to develop the models. All analyses were performed on a GPU machine with a 32-core AMD processor with 128 GB RAM (Advanced Micro Devices, Santa Clara, CA), 2 TB PCIe flash memory, 5 TB SDD Hard disks, and a single NVIDIA Titan V GPU with 12 GB VRAM.

**Results**

In this SEER database, the complete follow-up information was available for 1,941,893 cases. Here, 14.34\% of the patients died due to one of the urologic cancers, with an overall median follow-up of 12 years (range 1-44 years). Figure 1 illustrates the distribution of the cancer-specific death status over the year. When these patients stratified by the age at diagnosis, the age group (30-35 years) had a median follow up of 14 years, and patients belonging to the age group (65-69 years) had a median follow up of 13 years (Figure 1A). By stratifying according to the cancer-specific death status, the median time to cancer-
specific death was 11 years, and the median observation time was 13 years for those who were event-free at the time of completing the 18th version of the SEER database (Figure 1B and C).
Figure 1: (A) The median follow-up duration in years after stratifying by age at diagnosis (categorized); (B) the relative frequencies and (C) absolute frequencies for cancer-specific death status stratified by time from diagnosis (year).
Feature Importance

After model training, we evaluated the feature list after sorting by their importance from model exceeded concordance index 0.85. Although feature importance algorithms did weigh relevant features, the top features considered necessary by machine learning included inferior, obsolete, or redundant features from the clinical perspective or data acquisition. Consequently, the top feature list was not adequate from the clinical perspective.

Figure 2 illustrates the feature importance scores determined by (A) linear supportive machine (B) random forest classification (C) XGBoost classification (D) Simple recurrent neural network model (sRNN). We found weighing the features by machine learning not necessarily reflects the clinical importance of these features. (E) Shapely values for the sRNN model to prognosticate cancer-specific survival outcome. A high-resolution version is available as supplement file 2.
Table 1 lists the features selected for survival modeling. ** denotes that this feature was derived from other features. ++ the definition of the feature varies by the procedures to obtain prostate cancer tissues (e.g., biopsy or prostatectomy). Due to the variational options for cancer-specific features, we only provided the feature summary. Details regarding the available options can be obtained from https://www.myhealthadvisor.info.

CS: Cancer-specific; the number of CS site-specific factors is defined by the SEER program to differentiate between different CS site specific factors.

| Feature                              | Input data.                                      |
|--------------------------------------|--------------------------------------------------|
| **Common features**                  |                                                  |
| Age at diagnosis                     | Continuous value ranged between 30 and 114 years  |
| Race                                 | White                                            |
|                                      | Black                                            |
|                                      | Others                                           |
|                                      | Unknown                                          |
| Total number of malignant tumors for | Continuous value ranged between 1 and 10         |
| the patient                          |                                                  |
| Regional nodes positive 1988         | The number of positive lymph nodes:              |
|                                      | The range between 0 and 90                       |
|                                      | >=90                                             |
|                                      | Unknown or not performed (-1)                    |
| Regional nodes examined 1988         | The number of lymph nodes examined:              |
|                                      | The range between 0 and 90                       |
|                                      | >=90                                             |
|                                      | Unknown or not performed (-1)                    |
| SEER historic stage A **             | Tumor stage summary:                             |
|                                      | Localized                                        |
|                                      | Distant                                          |
|                                      | Regional                                         |
|                                      | Localized and/or regional (Prostate Cancer)      |
| Time, year                           | The follow-up duration until the last check or death. |
| Cancer-specific features | Prostate Cancer | Kidney Cancer | Urinary Bladder Cancer | Renal Pelvis or Ureter Cancer | Testicular Cancer | Penile Cancer |
|--------------------------|-----------------|---------------|------------------------|-----------------------------|-----------------|--------------|
| CS site-specific factor 2 | Thee measurement result of serum PSA level (ng/mL) | The vein involvement (Vena Cava and vena renalis) | The length of the maximum lymph node metastasis | Not applicable |
| CS Mets at dx 2004 | Evidence of metastases at the time of diagnosis: The definition varies between cancers of the genitourinary system. Detailed information for the cancer-specific list can be obtained from https://www.myhealthadvisor.info. |
| Derived AJCC M 6- ed 2004** | M0 | M0 | M0 | M0 | M1 | M1 |
| | MX | M1 | M1 | M1 | MX | MX |
| | M1 | MX | MX | MX | M1a | M1a |
| | M1b | M1b | M1b | M1c | MX | MX |
| CS site-specific factor 12 | Number of positive biopsy cores | Not applicable |
| CS site-specific factor 13 | Number of total biopsy cores | Not applicable | Postoperative Alpha Fetoprotein (AFP) level | Not applicable |
| CS extension 2004 | Tumor stage and extension: The definition varies between cancers of the genitourinary system. Detailed information for the cancer-specific list can be obtained from https://www.myhealthadvisor.info |
Model Development and Evaluation

Since the feature selection by machine learning was invalid, we manually selected 16 features for survival modeling as provided in Table 1. These features were covering the tumor stage with the description of the tumor extension, the status, and the condition of locoregional and distant metastases, the elevation status of the prostate-specific antigen (PSA) level at the time of diagnosis or the postoperative alpha-fetoprotein (AFP) level status, and the age at the diagnosis (Table 1). All urologists (n=5) have approved and agreed on the feature-list. Nine of these features were cancer specific. Since we intended to develop a single follow-up model for cancers of genitourinary organs, we fixed the input dimension to 16 features; not applicable features of a case were ignored by placing -1 into the corresponding column positions in the input data.

The Grid search analysis for hyperparameter configuration revealed that it is optimal to select the simple recurrent unit (sRNN) and RMSprop for the model designs. The Grid
search analysis did not recommend the inclusion of the dropout function. The model was trained on training set (n=1,747,703 cases, from which 250,555 cases died because of urologic cancers), and the validation set was used to optimize the model training. The test set had 194,190 cancer patients, from which 27,840 cases died due to urologic cancers. The overall concordance index on the test set was 0.80. The Kaplan–Meier (KM) curve revealed a well-fitted model for cancer-specific survival rates over a follow-up period of 35 years on the test set (Figure 3). When the test set is stratified by the cancer origins (organs), we found that the KM fitness mostly remained unchanged.
Figure 3 shows how well the predicted Kaplan–Meier (KM) estimates fit the observed KM estimates even after stratifying by cancers of different organs. A good overlap means a well-fitted model that facilitates the reconstruction of the unseen population’s survival history with cancer diseases.

The stratification by quantile thresholds showed well-discriminative risk groups (Log-rank P<0.001), as shown in Figure 4B. When we compare to the tumor dissemination status that used by the SEER program for survival estimation (Figure 4A), we identified that our model could introduce a very-low risk group with a 10-year cancer-specific survival (CSS) rate of 99.55% (95% Confidence interval -CI-: 99.43 – 99.64). In contrast, the best 10-year CSS rate of the tumor stages defined by the tumor dissemination status was for localized cancer diseases (CSS rate: 95.92%; 95% CI: 95.72 – 96.11). The 10-year CSS
rate of the distant cancer disease was 61.72% (95% CI: 60.75 – 62.67), whereas the 10-year CSS rate was 66.72% (95% CI: 65.49 – 66.88) for the high-risk group determined by the predicted probability.

Figure 4 (A) the Kaplan–Meier (KM) Curves for the SEER staging groups and (B) KM curve after stratifying the prognosticated survival probabilities using quantile thresholds (10%, 50%, 90%) defined on the training set. The KM curves form A and B were significantly discriminative (Log-rank P<0.005). For figure A, we calculated Log-rank P using localized, distant, regional, and regional/localized stage groups. Localized and regional prostate cancer cases were combined due to their high 5-year survival rates (nearly 100%) compared to other entities. The KM curves were generated on the test set.
Follow-up Management Solution

We prepared an access-controlled web portal that incorporates the AI solution ([https://www.myhealthadvisor.info](https://www.myhealthadvisor.info)) and covers 6 major urologic cancers for follow-up planning. This solution proposed intervals for follow-up and provided information about the cancer-specific follow-up score over the follow-up as a curve; it highlighted the case’s

---

1. A white man, 60 years old at time of diagnosis had PCa confirmed by biopsy that was indicated due to elevated PSA. The patient was treated with RPE (92') and had pT3bR0 pN1 (1/5) cM0 and Gleason grade 3 + 4 = 7a.

2. A white man, 65 years old at time of diagnosis had PCa confirmed by biopsy (cT1c) that was indicated due to elevated PSA of 5 ng/mL (08'). 2 of 12 biopsy cores were positive and the maximum tumor length was 15%. Gleason grade was 3 + 3. Active surveilance was recommended.

3. A white man, 67 years old at time of diagnosis had PCa confirmed by biopsy (1/10) that was indicated due to elevated PSA. The patient was treated with RPE (04') and had pT3aR0 pN1 (1/3) cM0 and Gleason grade 4 + 5 = 9.

4. A white man, 61 years old at time of diagnosis had PCa confirmed by biopsy that was indicated due to elevated PSA. Only the total number of biopsy cores is known (6). The patient was treated with RPE (95') and had pT3aR0 pN0 cM0 and Gleason grade 3 + 4 = 7a.

---

Figure 5 illustrates 4 cases with different clinical conditions and outcomes. We identified that the AI-based follow-up plan could capture all clinical events that led to a change in the treatment or the definition of the tumor condition. Another useful feature was the time suggestion to reconsider the follow-up plan. The red area denotes that the patients require a close surveillance; the red area is defined by the velocity of follow-up scores (See Material and Method for more details). The red line reveals a change in the velocity direction and suggests that the follow-up may need to be re-evaluated. The blue line or curve corresponds to the follow-up scores estimated by our model over the time from diagnosis. The black line corresponds to the recent position of the patient on the suggested surveillance plan. We identified that the curve course of the follow-up scores aligned with the treatment decision despite the model did not consider the treatment for the prognosis, emphasizing that the tumor stage inherits some information about the treatment definition and the potential cancer progression course.

PCa: Prostate Cancer, ADT: Androgen Deprivation Therapy; RTX: local radiation therapy by a local recurrence; BCR: biochemical recurrence; CRPC: Castration-Resistant Prostate Cancer which occurs when the cancer progresses despite ADT; PSA: Prostate-specific Antigen (ng/mL). 92' means 1992. The first case (1) received ADT directly after surgery (radical prostatectomy) in the first year of diagnosis due to the presence of lymph node metastases.
position on the follow-up score curve from the date of diagnosis; it also provided a probability range for the cancer-specific death. Further, the AI-solution suggested the number of visits per year to capture tumor progression and, in which year, we can omit the follow-up visits. We found that the follow-up suggestions were clinically meaningful for different clinical situations in four exemplary cases with prostate cancer not included in the SEER database, as Figure 5 illustrates.

**Discussion**

We successfully introduce an AI-assisted risk-adapted solution to manage the surveillance of patients with urologic cancer diseases. The follow-up management is based on the reconstruction of the cohort’s risk profile history from the national cancer registry (SEER). Our solution can suggest the annual visit number for follow up, the time point to consider modifying the follow-up plan, and when to apply a close follow-up plan. Further, our follow-up model is well-calibrated for the survival estimation over a long period, while providing stability and good fitness.

We identified that the observation intervals (red area in Figure 5) suggested by our algorithms covered the first 5 years mostly considered by the current clinical guidelines and the follow-up plan practices. Further, our approach has covered the time beyond 5 years where we do not have data from studies on the optimal observation management for urologic cancer diseases.

Our AI solution can deliver clinically meaningful follow-up plans as illustrated on men diagnosed with prostate cancer. For instance, our solution could correctly observe the example case who developed castration-resistant prostate cancer (CRPC).

Developing a follow-up model for cancer survivors is considered one of the major challenges in clinical routine. Various cancer organizations recommend using a risk-
stratified approach for care and transition for follow-up planning. Our solution may help the cancer patients get a data-driven follow-up plan personalized based on their risk profile.

The feature importance aims to explain the contribution of a feature to the prediction in a machine learning model, so the current study evaluated the feature importance provided by machine learning algorithms from the clinical perspective. Here, we identified that feature importance is not a reflection of clinical importance. Machine learning also tends to capture features that define the data patterns related to altered documentation behaviors. For instance, some machine learning approaches weighted outdated features in the current version of the database. Therefore, domain knowledge for the feature selection is essential in such large datasets to avoid having models biased toward outdated features or data patterns that are clinically irrelevant.

Given the variation in the follow-up coverage according to the age at diagnosis and the gap in the follow-up after 20 years (Figure 1), our follow-up management is reliable in the first 13 years after diagnosis. The results from the follow-up period between 13 and 22 years remain acceptable according to the comparison results between the predicted and true KM estimates. The results from the period beyond 22 years after diagnosis should be taken cautiously due to altered retrospective documentation procedures which are reflected in the gap in the follow-up data for the 23rd year and the unusual accumulation of death events or cases for 24th and 25th year in SEER database (Figure 1 and 3).

The current study has potential limitations or controversies that warrant mention. First, cancer progression is a dynamic process and influenced by different factors. However, our study approached a puzzle completion concept to develop survival modeling using a large dataset covering a long follow-up period to reduce this limitation. The diagnostic and therapeutic improvement has helped to detect cancers in the earlier stages, resulting
in longer survival as reflected in the SEER database. Most recent drug options for different urologic cancers found their first approval in early 2000. Although the SEER database does not include detailed treatment information, the version of the SEER database (1975-2017) considered by the current study very likely covered patient groups who received such approved drugs. There was a lack of information regarding patient comorbidities and complications after local treatment or information related to different treatment options, which undoubtedly influences progression and survival. The staging procedure was not standardized, which may have resulted in an underestimation/overestimation of the tumor extent and thereby influenced the data quality. However, the SEER-based data represent the actual situation in the daily clinical setting.

Further, the SEER database is the only comprehensive population-based database in the United States and represents an ideal approach for studying the survival of patients diagnosed with cancer diseases, especially in recent periods. We did not consider the treatment status in our survival modeling because the clinical and pathological information may include hidden information about the treatment strategy. Although we did not study other well-known prognostic factors like treatment response, we preferred the specific cancer death status as an endpoint to develop a follow-up model, because it is also a reflection of the treatment response and cancer progression risk. We successfully demonstrated that AI helps develop personalized follow-up strategies based on the risk profile in contrast to the current practice. We preferred the national cancer registry because it represents different clinical conditions of cancer diseases more than any study cohorts from a single institution or even from two institutions. We did not consider the tumor grading for bladder cancers due to the significant variation in the definition of tumor grading between the versions and the challenge of adapting the recent version of
tumor grading in clinical routine \cite{29,30}. We currently preferred AFP over the serum level of beta human chorionic gonadotropin (BHCG) for testicular cancer because AFP provides additional information on the subtype of testicular cancer compared to BHCG \cite{31}.

Prospectively, we aim to connect this AI solution to the Hospital Information System (HIS) using API (Application Programming Interface) and HL7 protocol as an additional tool to the HIS-based application for the real-time survival rate estimation introduced by our previous study \cite{32}. Further, another potential utilization of our AI-based follow-up planner is defining the follow-up for clinical trials.

In summary, the AI-based solution for follow-up management is feasible; it provides decision-aided and patient-friendly tools for a more personalized follow-up plan for men with prostate cancer. We believe that having a defined follow-up plan in advance provides better time management for the patients and healthcare personals. Further, the current study successfully unlocked the potential of national cancer registry to develop clinical-oriented solutions. Finally, we call the research teams from different geographies to validate our model independently as it is essential to validate it by different independent studies\cite{14}. A secured API access required to communicate with the solution will be provided for these research teams upon request. The API access will not be publicly available due to the security concerns.

**Acknowledgment**

We would like to thank the SEER program for providing this database.

**References**

1. Babjuk, M. *et al.* EAU guidelines on non–muscle-invasive urothelial carcinoma of the bladder: update 2016. *European urology* 71, 447-461 (2017).
2. Ljungberg, B. Renal cell carcinoma.
3. Rouprêt, M. *et al.* Urothelial carcinomas of the upper urinary tract.
Witjes, J. A. et al. Updated 2016 EAU guidelines on muscle-invasive and metastatic bladder cancer. *European urology* **71**, 462-475 (2017).

Stephenson, A. et al. Diagnosis and treatment of early stage testicular cancer: AUA guideline. *The journal of urology* **202**, 272-281 (2019).

Wilt, T. J. et al. Radical Prostatectomy or Observation for Clinically Localized Prostate Cancer: Extended Follow-up of the Prostate Cancer Intervention Versus Observation Trial (PIVOT). *European urology* (2020).

Wilt, T. J. et al. Follow-up of Prostatectomy versus Observation for Early Prostate Cancer. *N Engl J Med* **377**, 132-142, doi:10.1056/NEJMoai1615869 (2017).

Mottet, N. & Briers, E. EAU-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer.

Duggan, M. A., Anderson, W. F., Altekruse, S., Penberthy, L. & Sherman, M. E. The surveillance, epidemiology and end results (SEER) program and pathology: towards strengthening the critical relationship. *The American journal of surgical pathology* **40**, e94 (2016).

Pawade, D., Sakhapara, A., Jain, M., Jain, N. & Gada, K. Story scrambler-automatic text generation using word level rnn-lstm. (2018).

Hart, S. in *Game Theory* 210-216 (Springer, 1989).

Eminaga, O., Al-Hamad, O., Boegemann, M., Breil, B. & Semjonow, A. Combination possibility and deep learning model as clinical decision-aided approach for prostate cancer. *Health Informatics J*, 146045821955884, doi:10.1177/146045821955884 (2019).

Sutskever, I., Vinyals, O. & Le, Q. V. in *Advances in neural information processing systems*. 3104-3112.

Steyerberg, E. W. et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* **10**, e1001381, doi:10.1371/journal.pmed.1001381 (2013).

Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

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

Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Hinton, G., Srivastava, N. & Swersky, K. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent.

Statistics, N. C. f. H. *Life Expectancy*, [<https://www.cdc.gov/nchs/fastats/life-expectancy.htm>] (2020).

Harrell, F. E., Jr., Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* **15**, 361-387, doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4 (1996).

Royston, P. & Altman, D. G. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol* **13**, 33, doi:10.1186/1471-2288-13-33 (2013).

Betensky, R. A. Measures of follow-up in time-to-event studies: Why provide them and what should they be? *Clin Trials* **12**, 403-408, doi:10.1177/1740774515586176 (2015).

Tamada, S. et al. Time to progression to castration-resistant prostate cancer after commencing combined androgen blockade for advanced hormone-sensitive prostate cancer. *Oncotarget* **9**, 36966-36974, doi:10.18632/oncotarget.26426 (2018).

Jacobs, L. A. & Shulman, L. N. Follow-up care of cancer survivors: challenges and solutions. *Lancet Oncol* **18**, e19-e29, doi:10.1016/S1470-2247(16)30386-2 (2017).

Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *Journal of machine learning research* **3**, 1157-1182 (2003).

Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **68**, 394-424 (2018).

FDA. FDA expands approval of Sutent to reduce the risk of kidney cancer returning, [<https://www.fda.gov/news-events/press-announcements/fda-expands-approval-sutent-reduce-risk-kidney-cancer-returning>] (2017).

Oncology Drugs Approved by the FDA in Early 2018. *Am Health Drug Benefits* **11**, 209-210 (2018).

Boustead, G. B. et al. Stage, grade and pathological characteristics of bladder cancer in the UK: British Association of Urological Surgeons (BÂUS) urological tumour registry. *BJU Int* **113**, 924-930, doi:10.1111/bju.12468 (2014).

Grignon, D. J. The current classification of urothelial neoplasms. *Mod Pathol* **22 Suppl 2**, S60-69, doi:10.1038/modpathol.2008.235 (2009).
Germa-Lluch, J. R. et al. Clinical pattern and therapeutic results achieved in 1490 patients with germ-cell tumours of the testis: the experience of the Spanish Germ-Cell Cancer Group (GG). *Eur Urol* **42**, 553-562; discussion 562-553, doi:10.1016/s0302-2838(02)00439-6 (2002).

Dolin, R. H. et al. HL7 clinical document architecture, release 2. *Journal of the American Medical Informatics Association* **13**, 30-39 (2006).

Breil, B., Semjonow, A., Muller-Tidow, C., Fritz, F. & Dugas, M. HIS-based Kaplan-Meier plots--a single source approach for documenting and reusing routine survival information. *BMC Med Inform Decis Mak* **11**, 11, doi:10.1186/1472-6947-11-11 (2011).