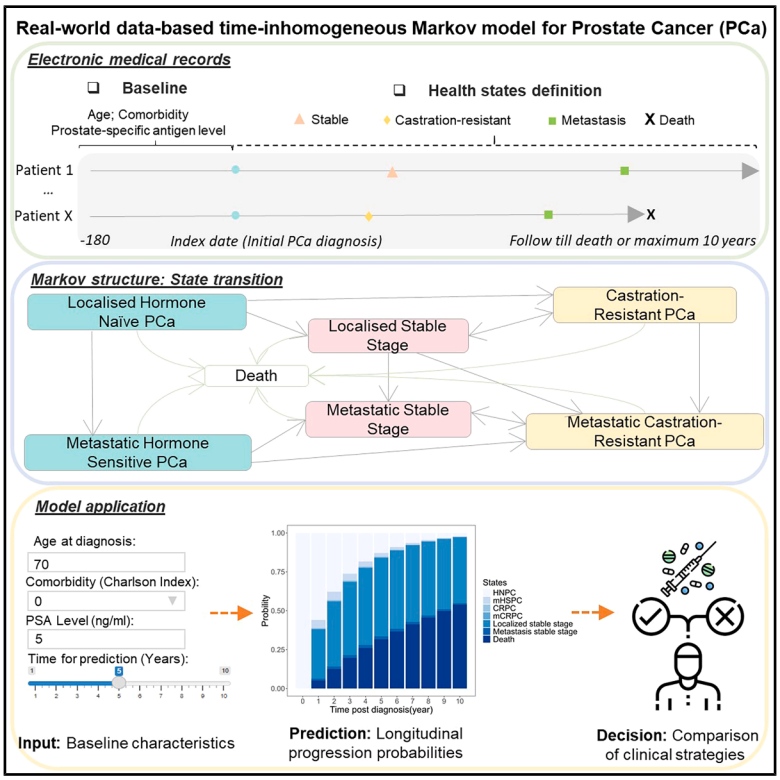
iScience

Article

Predicting disease progression and mortality in prostate cancer using real-world data-driven time-inhomogeneous Markov models

Graphical abstract



Highlights

- Time-inhomogeneous Markov model predicts multiple cancer progression endpoints
- Electronic medical record provides a rich data source to support complex model structures
- The model outperforms Cox and random survival forests for MFS and OS prediction

Authors

Jiaqi Wang, Yuanshi Jiao, Dawn Craig, ..., Qingpeng Zhang, Rong Na, Xue Li

Correspondence

yungna@hku.hk (R.N.), sxueli@hku.hk (X.L.)

In brief

Health sciences; Medicine; Medical specialty; Internal medicine; Oncology; Urology

Wang et al., 2026, iScience *29*, 117536
October 16, 2026
https://doi.org/10.1016/j.isci.2026.117536

CellPress

iScience

CellPress
OPEN ACCESS

Article

Predicting disease progression and mortality in prostate cancer using real-world data-driven time-inhomogeneous Markov models

Jiaqi Wang,[1,12] Yuanshi Jiao,[1,2,12] Dawn Craig,[3] Steven Wai Kwan Siu,[4] Lei Si,[5] David Makram Bishai,[6] Yi Yang,[7] Yingyao Chen,[7] Qingpeng Zhang,[8,10,11] Rong Na,[9,*] and Xue Li[1,2,13,*]

[1]Centre for Safe Medication Practice and Research, Department of Pharmacology and Pharmacy, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China
[2]Department of Medicine, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China
[3]Population Health Sciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK
[4]Department of Clinical Oncology, Queen Mary Hospital, Hong Kong SAR, China
[5]School of Health Sciences, Western Sydney University, Penrith, NSW, Australia
[6]Division of Health Economics, Policy and Management, School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China
[7]National Health Commission Key Laboratory of Health Technology Assessment, Fudan University, Shanghai, China
[8]HKU Musketeers Foundation Institute of Data Science, The University of Hong Kong, Hong Kong SAR, China
[9]Department of Surgery, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China
[10]Department of Pharmacology and Pharmacy, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China
[11]The University of Hong Kong-Shenzhen Hospital, Shenzhen, China
[12]These authors contributed equally
[13]Lead contact
*Correspondence: yungna@hku.hk (R.N.), sxueli@hku.hk (X.L.)
https://doi.org/10.1016/j.isci.2026.117536

SUMMARY

Prostate cancer progression varies across patients. Understanding disease trajectories and mortality risk is important for improved healthcare. This study developed a time-inhomogeneous Markov model using territory-wide electronic medical records from Hong Kong to estimate 10-year disease progression. The model was constructed using data from 3,274 patients newly diagnosed in 2010–2012. Their electronic medical records defined baseline covariates (age, Charlson Comorbidity Index [CCI], and prostate-specific antigen [PSA] level) and time-varying health states. Mild cases (age≤65, CCI = 0, PSA≤4 ng/mL) were predicted to have 16.9% and 28.0% mortality at five and ten years. Evaluated on an independent 2013 cohort, our model showed strong prediction performance for metastasis-free survival (concordance index [C-index]: 0.780; 95% confidence interval [CI]: 0.723–0.837) and overall survival (C-index: 0.787; 95% CI: 0.731–0.843), exceeding Cox proportional hazards and random survival forest models. This study provides a pragmatic tool for long-term progression risk prediction and health economic evaluation.

INTRODUCTION

Prostate cancer (PCa) is the most frequently diagnosed malignancy in males worldwide and a major health concern.[1] The incidence and mortality of PCa have increased substantially, with a rise of 116% and 108%, respectively, from 1990 to 2019, reaching 1.41 million new cases and 375,304 deaths in 2020.[2] As global population growth and aging rise, PCa presents an increasing economic burden.[2]

Understanding the progression of the disease and life expectancy enables timely and appropriate treatment and prevents unnecessary treatments for men who are unlikely to benefit. Various progression prediction tools have been developed for localized PCa, including the Cancer of the Prostate Risk Assessment (CAPRA) model[3] for biochemical recurrence and mortality risk prediction, Memorial Sloan Kettering Cancer Center (MSKCC) nomograms[4] and PREDICT Prostate[5] for post-treatment survival estimates. Time to metastasis is also a clinically significant outcome that has been explored in previous studies.[6,7] In clinical trials for localized PCa, metastasis-free survival (MFS) is validated to serve as a surrogate endpoint for overall survival (OS), given that metastatic disease is considered a lethal phenotype in PCa.[8] However, existing common survival models like the Cox-proportional hazard (Cox-PH) model can only analyze one outcome each time, for example, either metastasis or mortality. This limits their ability to provide a longitudinal and comprehensive prediction of disease trajectory.

The Markov-chain-based multistate model offers a more flexible approach by representing the disease's progression through

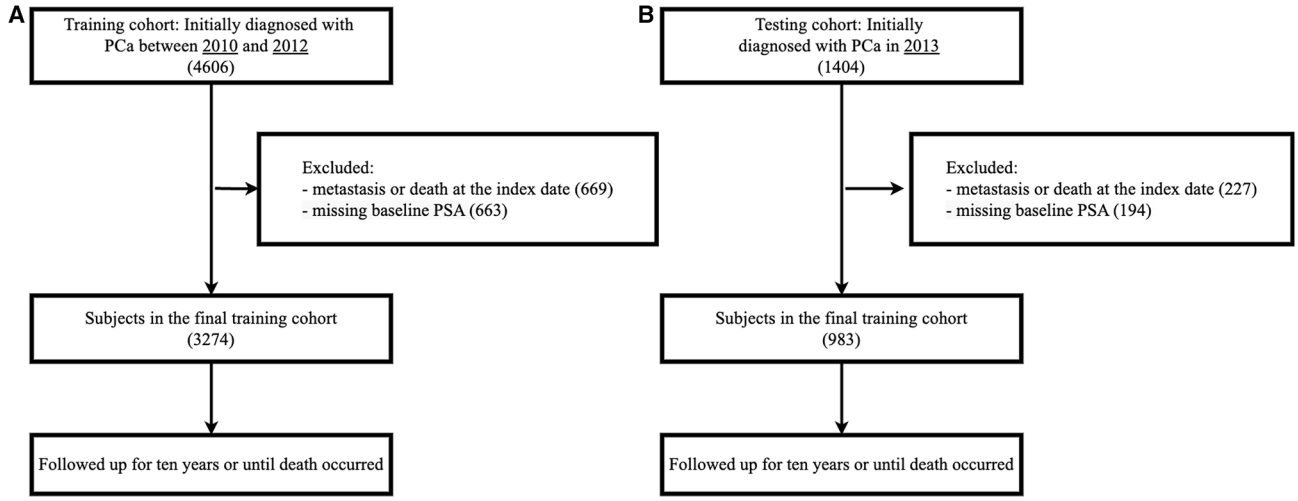


Figure 1. Flowchart of cohort filtering and follow-up

(A) Training cohort.

(B) Evaluation cohort.

PCa, prostate cancer; PSA, prostate-specific antigen.

health state transitions, enabling dynamic prediction of progression over time. For instance, Bhatt et al. used the multistate Markov model to depict the natural progression of PCa during the screening process, providing insight from an initial healthy state to screen detection, clinical diagnosis, and ultimately to death, which is challenging to capture with a simple Cox-PH model.[9] In the existing studies, estimating progression speed among states commonly relied on survival outcomes from clinical trials.[10,11] However, the utility of such multistate models might be constrained due to the limited follow-up time and the lack of individual-level information.[12,13] In contrast, electronic medical records (EMRs) offer a solution to these limitations by providing access to individual data across the full spectrum of disease states, enabling more comprehensive and realistic modeling of disease progression.

Existing Markov models of PCa have commonly adopted simplified three-state structures, distinguishing between the progression free, disease progression, and death.[14] While such models are straightforward and computationally efficient, they are insufficient to capture the full complexity of disease progression from initial diagnosis to clinically important endpoints, including time to castration-resistant[15,16] disease, metastasis, and death from any cause. Following the standard initial management of advanced PCa with androgen deprivation therapy, castration resistance typically develops within two to three years.[17] To distinguish treatment choice and intensity, major guidelines, such as the National Comprehensive Cancer Network and the European Association of Urology, structure advanced PCa progression based on castration resistance and metastasis status, including non-metastatic castration-resistant PCa, metastatic castration-sensitive PCa, and metastatic castration-resistant PCa.[18,19] Incorporating clinically relevant intermediate states, particularly metastasis and castration resistance, enables more realistic patient trajectories and more informative policy evaluations.[20]

In this study, we constructed a time-inhomogeneous Markov model to depict progression from PCa initial diagnosis to estimate 10-year progression probabilities. Clinically meaningful disease milestones were selected as health states to capture the key biological and prognostic transitions in PCa progression. Unlike conventional time-homogeneous models, which assume constant transition rates and identical sojourn times across all states, our framework incorporates both elapsed time since diagnosis and duration of time spent in each disease state.[21] The estimated progression was evaluated on different cohorts of patients to validate robustness.

RESULTS

Study population

Using a territory-wide EMR database managed by the Hospital Authority in Hong Kong, a total of 3,274 incident patients diagnosed with PCa between 2010 and 2012 (mean ± standard deviation [SD] age, 73.0 ± 8.59 years) were included in the training cohort for model development (Figure 1; Table 1). An additional 983 incident cases from 2013 (mean age: 73.0 ± 8.75 years) were reserved as an independent evaluation cohort (Figure 1; Table 1).

In the training cohort, 63.6% were free from medical conditions related to the Charlson Comorbidity Index (CCI) before the first PCa diagnosis. The median baseline prostate-specific antigen (PSA) level was 12.2 ng/mL (interquartile range [IQR]: 5.87–47.3 ng/mL), and 56.3% had a baseline PSA level greater than 10 ng/mL. Patients were followed until death or end of the study period (31 December 2023). During the follow-up, 1,530 (46.7%) deaths were recorded, with a median follow-up time of 3.97 years. Baseline characteristics, including age, PSA levels, and CCI, were well balanced between training and evaluation cohorts (Table 1).

Table 1. Baseline characteristics of the training and evaluation cohort

	Training cohort (*N* = 3274)	Evaluation cohort (*N* = 983)	*p* Value (if applicable)
Age (years)			
≤65	646 (19.7%)	207 (21.1%)	0.507
66∼70	548 (16.7%)	178 (18.1%)	
71∼75	777 (23.7%)	219 (22.3%)	
76∼80	678 (20.7%)	187 (19.0%)	
>80	625 (19.1%)	192 (19.5%)	
Charlson comorbidity index			
0	2083 (63.6%)	617 (62.8%)	0.155
1	758 (23.2%)	210 (21.4%)	
2	279 (8.5%)	104 (10.6%)	
≥3	154 (4.7%)	52 (5.3%)	
PSA (ng/mL)			
≤4	570 (17.4%)	147 (15.0%)	0.164
4.1–10	861 (26.3%)	257 (26.1%)	
>10	1843 (56.3%)	579 (58.9%)	
Metastasis			
no. of events	575 (17.6%)	158 (16.1%)	0.3
median follow-up time (years) [Min, Max]	1.37 [0.00548, 9.91]	1.46 [0.00274, 9.99]	0.554
Death			
no. of events	1530 (46.7%)	462 (47.0%)	0.912
median follow-up time (years) [Min, Max]	3.97 [0.00821, 10.0]	3.78 [0.00821, 9.98]	0.923
Medical history			
myocardial infarction	71 (2.2%)	19 (1.9%)	–
congestive heart failure	133 (4.1%)	31 (3.2%)	–
peripheral vascular disease	21 (0.6%)	8 (0.8%)	–
cerebrovascular disease	254 (7.8%)	80 (8.1%)	–
dementia	27 (0.8%)	10 (1.0%)	–
chronic pulmonary disease	217 (6.6%)	81 (8.2%)	–
rheumatoid disease	14 (0.4%)	2 (0.2%)	–
peptic ulcer disease	166 (5.1%)	52 (5.3%)	–
mild liver disease	91 (2.8%)	36 (3.7%)	–
diabetes without complications	292 (8.9%)	111 (11.3%)	–
diabetes with complications	53 (1.6%)	23 (2.3%)	–
hemiplegia or paraplegia	49 (1.5%)	9 (0.9%)	–
renal disease	73 (2.2%)	24 (2.4%)	–
malignant cancer	155 (4.7%)	60 (6.1%)	–
moderate or severe liver disease	4 (0.1%)	0 (0%)	–
metastatic solid tumor	4 (0.1%)	0 (0%)	–

PSA, prostate-specific antigen.

Disease progression prediction

Health states were defined based on international clinical guidelines and previous studies,[22–24] and identified using EMR (Table S1). Time-inhomogeneous multistate Markov models were then constructed to estimate PCa progression under Weibull distribution. In the main analysis, disease progression was categorized into four states: hormone-naive, localized progression, metastatic progression, and death (Figure 2A). In the sensitivity analysis, a more detailed seven-state model was specified, comprising hormone-naive, localized stable, localized castration resistant, metastatic hormone sensitive, metastatic stable, metastatic castration resistant, and death (Figure 2B). Baseline patient characteristics of age at diagnosis, CCI, and PSA level were incorporated as covariates to account for individual variability in disease trajectories.

Figure 3 illustrates the predicted progression trend under three hypothetical scenarios. Individuals with mild disease conditions at diagnosis (age ≤ 65 years old, CCI of 0, and PSA ≤ 4 ng/mL)

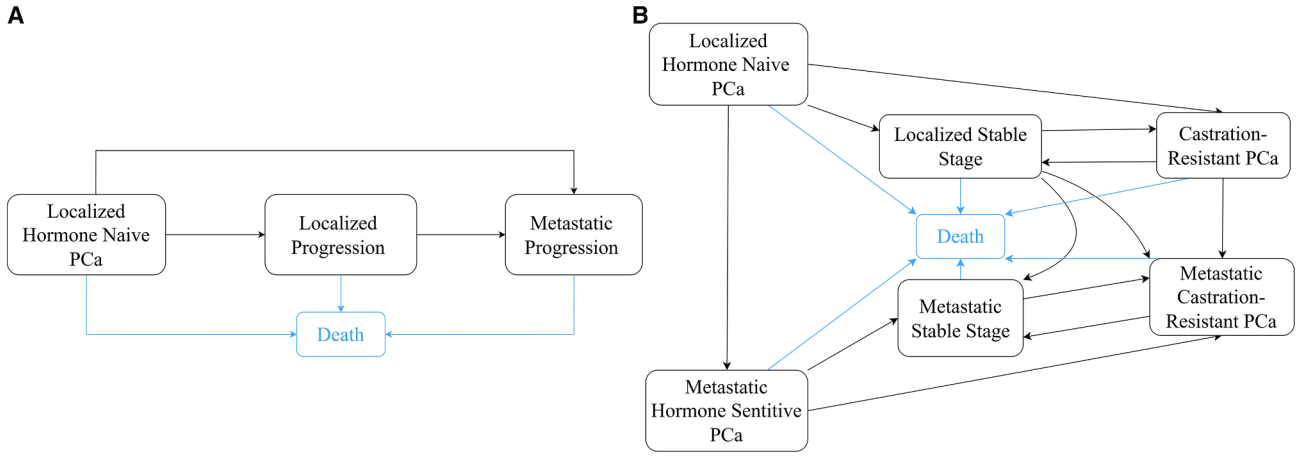


Figure 2. Transition dynamics for the Markov models

The diagrams show the allowable transitions between prostate cancer health states from initial diagnosis to death. Each box represents a disease state, and arrows indicate permitted transitions used in the time-inhomogeneous Markov models. Death is an absorbing state. Disease-state definitions are provided in Table S1.

(A) Four-state model structure used in the main analysis.

(B) Seven-state model structure used in the sensitivity analysis.

PCa, prostate cancer.

have a probability of 6.9% for developing metastatic cancer at year five post diagnosis; the cumulative mortality risk is 16.9% and 28.0% at year five and year ten from diagnosis. In comparison, severe patients (age >80 years old, CCI $\geq$ 3, and PSA level >10 ng/mL) showed lower metastasis progression probabilities (5.2% at year five; 2.6% at year ten), but higher mortality risks (40.4% and 58.6%, respectively, at year 5 and 10). Using proportions in the training cohort as weights, the weighted average probabilities of metastatic progression were 5.8% at five years and 3.7% at ten years, while the corresponding mortality probabilities were 21.7% and 35.0% at five and ten years, respectively (Table S2).

Evaluation

The predictive performance of the Markov models was evaluated for MFS and OS at ten years using the 2013 incident cohort. MFS was defined as the probability of remaining in non-metastatic states, and OS as the probability of not entering the absorbing death state. Both survival outcomes were derived from the estimated transition probabilities within the Markov framework. For comparison, Cox-PH and random survival forest (RSF) models were constructed using the same training cohort and baseline covariates (age, CCI, and PSA level). Model performance was subsequently compared with that of the time-inhomogeneous Markov model, using the concordance index (C-index) and the receiver operating characteristic (ROC) curve. Detailed model settings are provided in the STAR Methods sections.

OS prediction performance

The four-state time-inhomogeneous Markov model demonstrated strong OS prediction performance with a C-index of 0.787 (95% confidence interval [CI]: 0.731–0.843). This is higher than the Cox-PH model (0.607; 95% CI: 0.532–0.682) and RSF (0.589, 95% CI: 0.519–0.659). At the population level, the four-state model achieved an accuracy of 90.0% in predicting survival counts at five years (predicted 606 cases vs. observed 673) and 72.5% at ten years (predicted 366 vs. observed 505, Table S3). Sensitivity analysis shows the seven-state model exhibited an inferior OS prediction compared to the four-state model (C-index, 0.751; 95% CI: 0.690–0.811). Visualization through ROC curves further confirmed superior prediction of the four-state Markov model (Area Under the Curve [AUC], 0.787), followed by the seven-state model (AUC, 0.751), while Cox-PH (AUC, 0.607), and RSF (AUC, 0.589) performed less well (Figure S1).

MFS prediction performance

Evaluation of MFS yields a C-index of 0.780 (95% CI: 0.723–0.837). Comparative analyses revealed that the time-inhomogeneous Markov model outperformed both the Cox-PH model (C-index 0.630, 95% CI: 0.557–0.704) and RSF (C-index 0.617, 95% CI: 0.548–0.686). Longitudinal prediction accuracy for being event-free was maintained at an accuracy of 90.8% (predicted 640 cases vs. observed 705) at year five and 72.7% (predicted 379 cases vs. observed 521) at year ten (Table S3). Sensitivity analysis shows the seven-state model has a C-index of 0.748 (95% CI: 0.687–0.809). ROC results were consistent, with the four-state model performing best, followed by the seven-state model (AUC, 0.748), with lower performance observed for Cox-PH (AUC, 0.630) and RSF (AUC, 0.617) (Figure S1).

DISCUSSION

Accurate long-term progression prediction for patients with localized PCa is challenging but critical for effective patient counseling and treatment options.[25,26] In this study, we predicted the disease progression and all-cause mortality of PCa

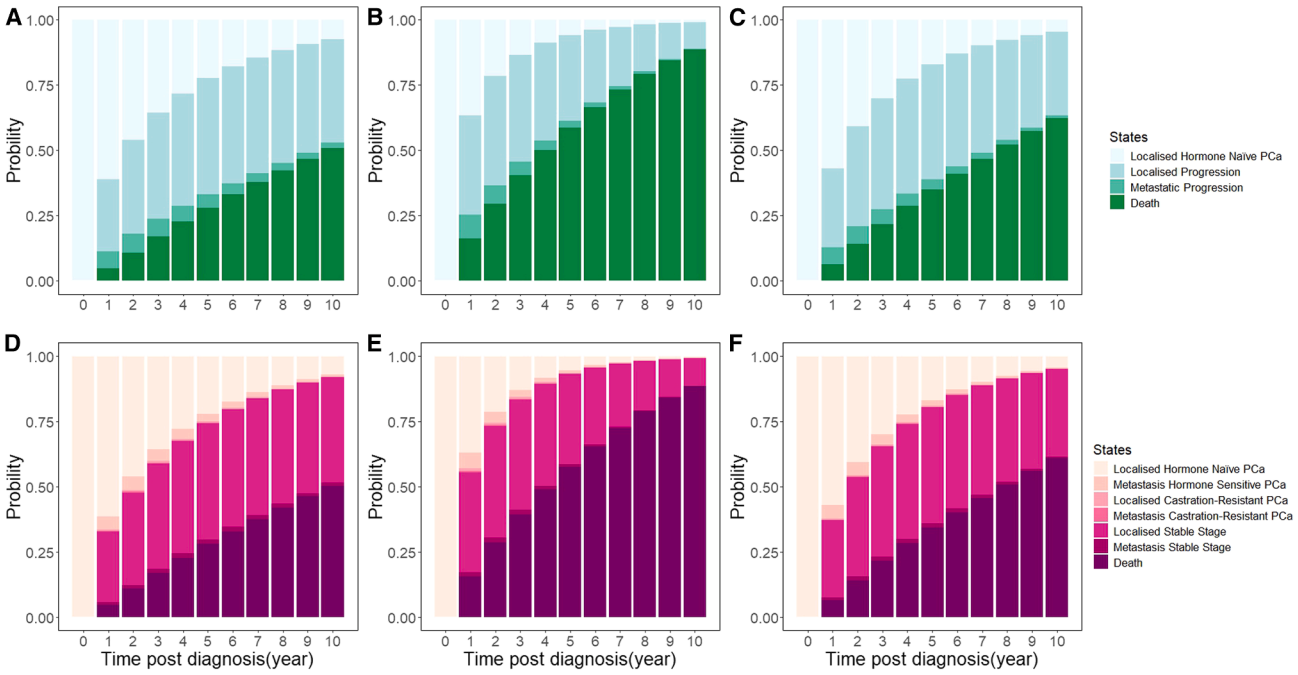


Figure 3. Predicted individual progression probabilities under three baseline scenarios

Stacked bars show the predicted probability of occupying each prostate cancer health state over 10 years after diagnosis. (A)–(C) show predictions from the four-state model used in the main analysis; (D)–(F) show predictions from the seven-state model used in the sensitivity analysis. For each model, probabilities are presented under three baseline scenarios. (A) and (D), mild condition: age ≤65 at initial diagnosis, baseline Charlson Comorbidity Index (CCI) of 0, and prostate-specific antigen (PSA) levels ≤4 ng/mL; (B) and (E), severe condition: baseline age >80 years old, CCI level ≥3, and PSA greater than 10 ng/mL; (C) and (F), weighted average: the averaged progression probabilities with weight calculated based on the proportion of patients under joint levels of covariations in the training cohort. PCa, prostate cancer.

patients through the time-inhomogeneous Markov model. By incorporating baseline covariates, we accounted for variations in progression across disease trajectories, spanning initial diagnosis to metastatic cancer and death. The C-index suggests the prediction for MFS and OS outperforms the classical Cox model and machine learning methods.

Through evaluation in an independent cohort, our model demonstrates good prediction accuracy for the ten-year all-cause mortality prediction (C-index, 0.787; 95% CI: 0.731–0.843), which was close to the reported top-ranked models, such as PREDICT Prostate (C-index, 0.79; 95% CI: 0.79–0.80) and CAPRA score (C-index varies between studies from 0.665 to 0.7).[3,27] Notably, this performance was achieved even without information on the biopsy and clinical stage, highlighting the model's ability to provide reliable prognostic prediction using more readily available data, making the model results highly practical for clinical use.

The inhomogeneous time setting employed in our model addresses the common limitation of the Markovian assumption, that future states depend solely on the current state and ignore historical information.[28] By incorporating time since initial diagnosis, the model captures temporal dynamics more effectively, which may contribute to its improved predictive performance.

The multistate modeling approach provides interpretable progression pathways for PCa, offering richer clinical insights than conventional single-endpoint prediction models. By explicitly modeling longitudinal transitions between health states, it enables dynamic assessment of patient trajectories over time and generates individualized estimates of metastasis and mortality risk, which can support clinical risk stratification and patient counseling.[29] Beyond prediction, the framework enables simulation of disease progression under uncertainty, thereby facilitating the evaluation and comparison of treatment strategies, and informing clinical decision-making and health policy planning.[30]

Applying EMR data and measurable real-world indicators now permits a more refined and realistic depiction of disease progression in clinical settings, incorporating details and complexity, with health states structured to reflect disease severity, treatment status, or risk or prognosis groups.[10,31] However, additional model complexity does not necessarily lead to improved performance. The sensitivity analysis of this study shows marginally lower C-index values in the seven-state model for both MFS (four-state model vs. seven-state model, 0.780 vs. 0.748) and OS (four-state model vs. seven-state model, 0.787 vs. 0.751). This finding underscores the importance of the trade-off between avoiding complexity and accuracy in multi-state models. Therefore, model development should aim for a parsimonious structure, balancing simplicity and accuracy by including only those states that meaningfully contribute to the prediction goal. In practice, this means constructing a model that is as simple as possible but sufficiently detailed to address

the research question, with the decision purpose guiding the number of health states.

In conclusion, this study provides a practical tool for predicting PCa progression at ten years. The multistate Markov model structure and estimated time-inhomogeneous transition probabilities provide an interpretable approach for characterizing disease trajectories and may support future applications in disease progression prediction and health decision-making.

Limitations of the study

Several limitations should be acknowledged, largely arising from the structured nature of the EMR database. First, this study is limited by the lack of pathology and histology information to confirm the tumor stage.[32] Some aspects of disease severity are partially captured through available measures: PSA levels correlate with tumor burden,[33] while the Charlson comorbidity index reflects competing risks of mortality.[34] Nevertheless, the absence of pathology and histology variables may reduce model precision and limit the ability to fully adjust for disease severity.[8] Second, time-varying covariates, including subsequent treatments, were not explicitly incorporated into the risk prediction model but were partially represented through the identification of health states. For example, transitions to the castration-resistant state represent a key clinical milestone reflecting the development of resistance to androgen deprivation therapy. This state was defined using time-varying clinically relevant criteria, including PSA levels and testosterone measurements. The model assumes that patients in a well-developed healthcare system generally receive standard-of-care treatment, but deviations from this assumption may lead to misestimation of transition risks and limit generalizability to other settings. Third, the study is based on data from a single healthcare system. Although the dataset includes a large, comprehensive set of anonymized electronic medical records from public hospitals in Hong Kong, the findings may not be directly generalizable to settings with different patient ethnicities, screening practices, treatment patterns, and healthcare access.

Future studies could externally validate these findings in diverse populations and healthcare systems. In addition, the model could be further enhanced by incorporating richer clinical and pathological data to better capture treatment heterogeneity and causal effects. Integration of state-of-the-art machine learning or deep learning algorithms[35] within the multistate modeling framework could enhance predictive accuracy without compromising the interpretability of disease progression.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Xue Li, (sxueli@hku.hk).

Materials availability

No reagents were generated in this study.

Data and code availability

- Data: Individual-level data cannot be publicly shared because of privacy and confidentiality requirements. Data were obtained from the Hospital Authority of Hong Kong and may be accessed for research purposes through the Hospital Authority Data Sharing Portal (https://www3.ha.org.hk/data).
- Code: The analysis code used in this study is publicly available through the SCAN2030 GitHub repository: https://github.com/scan2030/WP-2_PC-Mortality-Prediction.
- Other items: Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon reasonable request.

ACKNOWLEDGMENTS

This research was supported by the Hong Kong University Grants Committee Research Impact Fund (reference number: R7007-22), Shenzhen Medical Research Fund of the Shenzhen Medical Academy of Research and Translation (No. B032608), the General Research Fund of the Research Grants Council of Hong Kong (No. 17209225, No. 17210226), and the Seed Fund for Collaborative Research of The University of Hong Kong (No. 2407102490).

AUTHOR CONTRIBUTIONS

J.W., data curation, formal analysis, visualization, and writing – original draft; Y.J., data curation, formal analysis, and writing – review and editing; D.C., writing – review and editing; S.W.K.S., writing – review and editing; L.S., writing – review and editing; D.M.B., writing – review and editing; Y.Y., writing – review and editing; Y.C., writing – review and editing; Q.Z., writing – review and editing; R.N., supervision, methodology, and writing – review and editing; X.L., funding acquisition, conceptualization, supervision, methodology, and writing – review and editing.

DECLARATION OF INTERESTS

X.L. received research grants or contracts from the Health and Medical Research Fund (HMRF) Main Scheme and HMRF Fellowship Scheme of the Hong Kong Special Administrative Region (SAR), China, and from the Research Grants Council Early Career Scheme, Research Impact Fund (Hong Kong SAR); is also the former nonexecutive director of ADAMS Hong Kong SAR, China; received commission grants from Department of Health, Government of Hong Kong SAR, China; commission grants from Hospital Authority of Hong Kong SAR, China, internal funding from the University of Hong Kong, Hong Kong SAR, China, and research or education grants from Pfizer, Janssen and Bristol Myers Squibb (BMS), Novartis; received consultancy fees from Merck Sharp & Dohme, Pfizer, Open Health and The Office of Health Economics; and received honoraria for associate editorship from Nature Springer.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Data source
  - Patients and follow-up
  - Ethics statement
  - Health state definition and Markov structure
  - Covariates
  - Time-inhomogeneous Markov model
  - Model evaluation
  - Model application
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2026.117536.

Received: October 14, 2025
Revised: June 3, 2026
Accepted: August 31, 2026

REFERENCES

1. Leslie, S.W., Soon-Sutton, T.L., Sajjad, H., and Skelton, W.P. (2024). Prostate Cancer. In StatPearls (StatPearls Publishing LLC).
2. Zhang, W., Cao, G., Wu, F., Wang, Y., Liu, Z., Hu, H., and Xu, K. (2023). Global Burden of Prostate Cancer and Association with Socioeconomic Status, 1990-2019: A Systematic Analysis from the Global Burden of Disease Study. J. Epidemiol. Glob. Health *13*, 407–421. https://doi.org/10.1007/s44197-023-00103-6.
3. Brajtbord, J.S., Leapman, M.S., and Cooperberg, M.R. (2017). The CAPRA score at 10 years: contemporary perspectives and analysis of supporting studies. Eur. Urol. *71*, 705–709. https://doi.org/10.1016/j.eururo.2016.08.065.
4. Center TMSKC (2025). Prostate Cancer Nomograms. https://www.mskcc.org/nomograms/prostate.
5. Devos, G., and Joniau, S. (2020). PREDICT Prostate, a useful tool in men with low- and intermediate-risk prostate cancer who are hesitant between conservative management and active treatment. BMC Med. *18*, 213. https://doi.org/10.1186/s12916-020-01681-z.
6. Thomsen, F.B., Westerberg, M., Garmo, H., Robinson, D., Holmberg, L., Ulmert, H.D., and Stattin, P. (2020). Prediction of metastatic prostate cancer by prostate-specific antigen in combination with T stage and Gleason Grade: Nationwide, population-based register study. PLoS One *15*, e0228447. https://doi.org/10.1371/journal.pone.0228447.
7. Ladner, L.R., Adhikari, S., Bhutada, A.S., Cuoco, J.A., Patel, V.M., Entwistle, J.J., Rogers, C.M., and Marvin, E.A. (2024). Traditional Prostate Cancer Risk Assessment Scales Do Not Predict Outcomes from Brain Metastases: A Population-Based Predictive Nomogram. Cancers (Basel) *16*, 3029. https://doi.org/10.3390/cancers16173029.
8. San Francisco, I.F., Rojas, P.A., Bravo, J.C., Díaz, J., Ebel, L., Urrutia, S., Prieto, B., and Cerda-Infante, J. (2023). Can We Predict Prostate Cancer Metastasis Based on Biomarkers? Where Are We Now? Int. J. Mol. Sci. *24*, 12508. https://doi.org/10.3390/ijms241512508.
9. Bhatt, R., van den Hout, A., and Pashayan, N. (2021). A multistate survival model of the natural history of cancer using data from screened and unscreened population. Stat. Med. *40*, 3791–3807. https://doi.org/10.1002/sim.8998.
10. Therneau, T.M., and Ou, F.S. (2024). Using multistate models with clinical trial data for a deeper understanding of complex disease processes. Clin. Trials *21*, 531–540. https://doi.org/10.1177/17407745241267862.
11. Barbier, M.C., Tomonaga, Y., Menges, D., Yebyo, H.G., Haile, S.R., Puhan, M.A., and Schwenkglenks, M. (2022). Survival modelling and cost-effectiveness analysis of treatments for newly diagnosed metastatic hormone-sensitive prostate cancer. PLoS One *17*, e0277282. https://doi.org/10.1371/journal.pone.0277282.
12. Sathianathen, N.J., Alarid-Escudero, F., Kuntz, K.M., Lawrentschuk, N., Bolton, D.M., Murphy, D.G., Kim, S.P., and Konety, B.R. (2019). A Cost-effectiveness Analysis of Systemic Therapy for Metastatic Hormone-sensitive Prostate Cancer. Eur. Urol. Oncol. *2*, 649–655. https://doi.org/10.1016/j.euo.2019.01.004.
13. Zhang, P.F., Xie, D., and Li, Q. (2021). Adding Enzalutamide to First-Line Treatment for Metastatic Hormone-Sensitive Prostate Cancer: A Cost-Effectiveness Analysis. Front. Public Health *9*, 608375. https://doi.org/10.3389/fpubh.2021.608375.
14. Nguyen, T.T., Ameyaw, D., Obeng, G.D., Amuah, R., Józwiak-Hagymásy, J., Dóczi, T., Mezei, D., Németh, B., Tordai, A., Alanya, A., et al. (2025). Systematic Literature Review on Economic Evaluations and Health Economic Models in Metastatic Castration-Sensitive Prostate Cancer. Curr. Oncol. *32*, 412. https://doi.org/10.3390/curroncol32080412.
15. Scher, H.I., Morris, M.J., Stadler, W.M., Higano, C., Basch, E., Fizazi, K., Antonarakis, E.S., Beer, T.M., Carducci, M.A., Chi, K.N., et al. (2016). Trial Design and Objectives for Castration-Resistant Prostate Cancer: Updated Recommendations From the Prostate Cancer Clinical Trials Working Group 3. J. Clin. Oncol. *34*, 1402–1418. https://doi.org/10.1200/JCO.2015.64.2702.
16. Scher, H.I., Morris, M.J., Basch, E., and Heller, G. (2011). End points and outcomes in castration-resistant prostate cancer: from clinical trials to clinical practice. J. Clin. Oncol. *29*, 3695–3704. https://doi.org/10.1200/JCO.2011.35.8648.
17. Chandrasekar, T., Yang, J.C., Gao, A.C., and Evans, C.P. (2015). Mechanisms of resistance in castration-resistant prostate cancer (CRPC). Transl. Androl. Urol. *4*, 365–380. https://doi.org/10.3978/j.issn.2223-4683.2015.05.02.
18. Spratt, D.E., Srinivas, S., Adra, N., Ahmed, B., An, Y., Bitting, R., Chapin, B., Cheng, H.H., Cho, S.Y., D'Amico, A.V., et al. (2025). Prostate Cancer, Version 3.2026, NCCN Clinical Practice Guidelines In Oncology. J. Natl. Compr. Canc. Netw. *23*, 469–493. https://doi.org/10.6004/jnccn.2025.0052.
19. Tilki, D., van den Bergh, R.C.N., Briers, E., Van den Broeck, T., Brunckhorst, O., Darraugh, J., Eberli, D., De Meerleer, G., De Santis, M., Farolfi, A., et al. (2024). EAU-EANM-ESTRO-ESUR-ISUP-SIOG Guidelines on Prostate Cancer. Part II-2024 Update: Treatment of Relapsing and Metastatic Prostate Cancer. Eur. Urol. *86*, 164–182. https://doi.org/10.1016/j.eururo.2024.04.010.
20. Verry, C., Vincendeau, S., Massetti, M., Blachier, M., Vimont, A., Bazil, M.L., Bernardini, P., Pettré, S., and Timsit, M.O. (2022). Pattern of Clinical Progression Until Metastatic Castration-Resistant Prostate Cancer: An Epidemiological Study from the European Prostate Cancer Registry. Target. Oncol. *17*, 441–451. https://doi.org/10.1007/s11523-022-00899-6.
21. Cheung, L.C., Albert, P.S., Das, S., and Cook, R.J. (2022). Multistate models for the natural history of cancer progression. Br. J. Cancer *127*, 1279–1288. https://doi.org/10.1038/s41416-022-01904-5.
22. Pham, T., Sadowski, M.C., Li, H., Richard, D.J., d'Emden, M.C., and Richard, K. (2015). Advances in hormonal therapies for hormone naïve and castration-resistant prostate cancers with or without previous chemotherapy. Exp. Hematol. Oncol. *5*, 1–11.
23. Parker, C., Castro, E., Fizazi, K., Heidenreich, A., Ost, P., Procopio, G., Tombal, B., and Gillessen, S.; ESMO Guidelines Committee (2020). Prostate cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann. Oncol. *31*, 1119–1134. https://doi.org/10.1016/j.annonc.2020.06.011.
24. Cornford, P., van den Bergh, R.C.N., Briers, E., Van den Broeck, T., Cumberbatch, M.G., De Santis, M., Fanti, S., Fossati, N., Gandaglia, G., Gillessen, S., et al. (2021). EAU-EANM-ESTRO-ESUR-SIOG guidelines on prostate cancer. Part II—2020 update: treatment of relapsing and metastatic prostate cancer. Eur. Urol. *79*, 263–282. https://doi.org/10.1016/j.eururo.2020.09.046.
25. Saito, S., Sakamoto, S., Higuchi, K., Sato, K., Zhao, X., Wakai, K., Kanesaka, M., Kamada, S., Takeuchi, N., Sazuka, T., et al. (2023). Machine-learning predicts time-series prognosis factors in metastatic prostate cancer patients treated with androgen deprivation therapy. Sci. Rep. *13*, 6325. https://doi.org/10.1038/s41598-023-32987-6.
26. Kerkmeijer, L.G.W., Monninkhof, E.M., van Oort, I.M., van der Poel, H.G., de Meerleer, G., and van Vulpen, M. (2016). PREDICT: model for prediction of survival in localized prostate cancer. World J. Urol. *34*, 789–795. https://doi.org/10.1007/s00345-015-1691-4.
27. Cooperberg, M.R., Broering, J.M., and Carroll, P.R. (2009). Risk assessment for prostate cancer metastasis and mortality at the time of diagnosis. J. Natl. Cancer Inst. *101*, 878–887. https://doi.org/10.1093/jnci/djp122.
28. Titman, A.C., and Putter, H. (2022). General tests of the Markov property in multi-state models. Biostatistics *23*, 380–396. https://doi.org/10.1093/biostatistics/kxaa030.
29. Taguchi, A., Hara, K., Tomio, J., Kawana, K., Tanaka, T., Baba, S., Kawata, A., Eguchi, S., Tsuruga, T., Mori, M., et al. (2020). Multistate Markov Model

to Predict the Prognosis of High-Risk Human Papillomavirus-Related Cervical Lesions. Cancers (Basel) *12*, 270. https://doi.org/10.3390/cancers12020270.

30. Tang, J., Liang, Y., O'Neill, C., Kee, F., Jiang, J., and Congdon, N. (2019). Cost-effectiveness and cost-utility of population-based glaucoma screening in China: a decision-analytic Markov model. Lancet. Glob. Health *7*, e968–e978. https://doi.org/10.1016/S2214-109X(19)30201-3.
31. Le-Rademacher, J.G., Therneau, T.M., and Ou, F.S. (2022). The Utility of Multistate Models: A Flexible Framework for Time-to-Event Data. Curr. Epidemiol. Rep. *9*, 183–189. https://doi.org/10.1007/s40471-022-00291-y.
32. Barsouk, A., Padala, S.A., Vakiti, A., Mohammed, A., Saginala, K., Thandra, K.C., Rawla, P., and Barsouk, A. (2020). Epidemiology, Staging and Management of Prostate Cancer. Med. Sci. *8*, 28. https://doi.org/10.3390/medsci8030028.
33. Mandel, P., Hoeh, B., Preisser, F., Wenzel, M., Humke, C., Welte, M.N., Jerrentrup, I., Köllermann, J., Wild, P., Tilki, D., et al. (2021). Influence of Tumor Burden on Serum Prostate-Specific Antigen in Prostate Cancer Patients Undergoing Radical Prostatectomy. Front. Oncol. *11*, 656444. https://doi.org/10.3389/fonc.2021.656444.
34. Santos, A., Mattiolli, A., Carvalheira, J.B., Ferreira, U., Camacho, M., Silva, C., Costa, F., Matheus, W., Lima, M., and Etchebehere, E. (2021). PSMA whole-body tumor burden in primary staging and biochemical recurrence of prostate cancer. Eur. J. Nucl. Med. Mol. Imaging *48*, 493–500. https://doi.org/10.1007/s00259-020-04981-x.
35. Lee, C., Light, A., Alaa, A., Thurtle, D., van der Schaar, M., and Gnanapragasam, V.J. (2021). 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. Lancet Digit. Health *3*, e158–e165. https://doi.org/10.1016/S2589-7500(20)30314-9.
36. Chan, V.K.Y., Leung, M.Y.M., Chan, S.S.M., Yang, D., Knapp, M., Luo, H., Craig, D., Chen, Y., Bishai, D.M., Wong, G.H.Y., et al. (2024). Projecting the 10-year costs of care and mortality burden of depression until 2032: a Markov modelling study developed from real-world data. Lancet Reg. Health West. Pac. *45*, 101026. https://doi.org/10.1016/j.lanwpc.2024.101026.
37. Adeoye, J., Choi, S.W., and Thomson, P. (2020). Bayesian disease mapping and the 'High-Risk' oral cancer population in Hong Kong. J. Oral Pathol. Med. *49*, 907–913. https://doi.org/10.1111/jop.13045.
38. Au, P.C., Lee, A.W., Lee, V.H., Wong, I.C., Hui, R.Y., and Cheung, C.L. (2024). The trends in lung cancer prevalence, incidence, and survival in Hong Kong over the past two decades (2002-2021): a population-based study. Lancet Reg. Health West. Pac. *45*, 101030. https://doi.org/10.1016/j.lanwpc.2024.101030.
39. Jiao, Y., Ho, I., Li, T., Na, R., Wong, C., Wang, J., Siu, S.W.K., Wei, Y., Chen, Y., Chan, E.W., and Li, X. (2025). Real-world effectiveness of novel hormonal agents and docetaxel in patients with prostate cancer: A head-to-head comparison. iScience *28*, 112249. https://doi.org/10.1016/j.isci.2025.112249.
40. Shami, J.J.P., Yan, V.K.C., Wei, Y., Alwafi, H., Blais, J.E., Wan, E., Wong, C.K.H., Cheung, K.S., Leung, W.K., Wong, M.C.S., et al. (2023). Low-dose aspirin does not lower the risk of colorectal cancer in patients with type 2 diabetes taking metformin. J. Intern. Med. *293*, 371–383. https://doi.org/10.1111/joim.13590.
41. Wang, W., Zhang, Q., Thomson, P., Sharma, D., Ramamurthy, P., and Choi, S.W. (2023). Predicting oral cancer survival-Development and validation of an Asia-Pacific nomogram. J. Oral Pathol. Med. *52*, 628–636. https://doi.org/10.1111/jop.13454.
42. Yu, S.Y., Ip, M.S.M., Li, X., Cheung, K.S., Ren, Q.W., Wu, M.Z., Li, H.L., Wong, P.F., Tse, H.F., and Yiu, K.H. (2022). Low-dose aspirin and incidence of lung carcinoma in patients with chronic obstructive pulmonary disease in Hong Kong: A cohort study. PLoS Med. *19*, e1003880. https://doi.org/10.1371/journal.pmed.1003880.
43. Albertsen, P.C., Moore, D.F., Shih, W., Lin, Y., Li, H., and Lu-Yao, G.L. (2011). Impact of comorbidity on survival among men with localized prostate cancer. J. Clin. Oncol. *29*, 1335–1341. https://doi.org/10.1200/JCO.2010.31.2330.
44. Scosyrev, E., Messing, E.M., Mohile, S., Golijanin, D., and Wu, G. (2012). Prostate cancer in the elderly: frequency of advanced disease at presentation and disease-specific mortality. Cancer *118*, 3062–3070. https://doi.org/10.1002/cncr.26392.
45. Rajan, P., Sooriakumaran, P., Nyberg, T., Akre, O., Carlsson, S., Egevad, L., Steineck, G., and Wiklund, N.P. (2017). Effect of Comorbidity on Prostate Cancer-Specific Mortality: A Prospective Observational Study. J. Clin. Oncol. *35*, 3566–3574. https://doi.org/10.1200/JCO.2016.70.7794.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Territory-wide electronic medical records	Hospital Authority of Hong Kong	Data Sharing Portal: https://www3.ha.org.hk/data.
Software and algorithms		
R, version 4.2.2.	The R Foundation	https://www.R-project.org/
flexsurv: Flexible Parametric Survival and Multi-State Models (R package)	The Comprehensive R Archive Network	https://github.com/chjackson/flexsurv
msm: Multi-State Markov and Hidden Markov Models in Continuous Time (R package)	The Comprehensive R Archive Network	https://github.com/chjackson/msm
mstate: Data Preparation, Estimation and Prediction in Multi-State Models (R package)	The Comprehensive R Archive Network	https://github.com/hputter/mstate

METHOD DETAILS

Data source

We constructed Markov models based on a territory-wide EMR database managed by the Hospital Authority in Hong Kong. This dataset encompasses 13 million EMRs in all public hospitals in Hong Kong and accounts for approximately 73% of total hospital admissions.[36] It includes identity-anonymized records of diagnoses, prescriptions, death records, laboratory test results, and more. It has been extensively utilized in previous oncology-related clinical and epidemiological studies.[37–42]

Patients and follow-up

Incident patients who were initially diagnosed with PCa (International Classification of Diseases -Ninth Revision- Clinical Modification [ICD-9-CM]: 185) between 2010 and 2012 were included in our study's training cohort. Incident cases for 2013 were reserved as a validation cohort. The exclusion criteria are: 1) confirmed metastasis or death on the first day of diagnosis; 2) incomplete documentation of baseline Prostate-Specific Antigen (PSA), which is defined as laboratory test results within 180 days before the initial diagnosis. Patients were followed up until death or the end of the study period (31 December, 2023), whichever occurred first. With a maximum follow-up of ten years, the training cohort was used to construct the Markov model and estimate the transition probabilities.

Ethics statement

The study was approved by the Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (Institutional Review Board [IRB] reference number: UW 22-279). Given the completely de-identified nature of the EMR data utilized from the Hospital Authority database, patient consent was deemed unnecessary for this study.

Health state definition and Markov structure

Based on clinical definitions from international guidelines and prior studies,[22–24] we identified markers of disease states from the EMRs using diagnosis and laboratory test results. Metastatic progression was identified through the diagnosis of a subsequent malignant neoplasm after PCa, using ICD-9-CM codes 196–199. Castration resistance was determined following the European Association of Urology guidelines, based on consecutive increases in PSA despite castrate testosterone levels.[24] Additionally, we defined remission of PCa as a period of continuously low PSA levels. In the main analysis, we employed a simplified framework (Figure 2A) that categorized cancer progression into four states: hormone naive, localized progression, metastatic progression, and death.

To evaluate whether increasing the number of health states enhances predictive accuracy, we also constructed a more detailed seven-state Markov model in the sensitivity analysis (Figure 2B). In this model, disease begins with an initial state of hormone-naïve PCa, from which the disease may progress to either metastatic PCa or castration-resistant PCa, with all disease trajectories ultimately leading to death. A detailed definition of each state can be found in Table S1.

Covariates

In this study, we incorporated three baseline covariates: age at diagnosis, baseline CCI,[43] and baseline PSA level. Age at diagnosis was categorized into younger than 65, older than 80, and five-year intervals in between. The baseline CCI was determined through

medical history at the initial diagnosis of PCa. In addition, we defined the baseline PSA level as the averaged measurement obtained within 180 days before diagnosis and categorized this into three levels: less than or equal to 4 ng/mL, 4.1 to 10 ng/mL, and above 10 ng/mL.[44,45]

Time-inhomogeneous Markov model

A parametric time-inhomogeneous Markov model was constructed in this study to estimate the progression rate of PCa, with transition probabilities modeled as functions of time since diagnosis and baseline covariates. A clock-forward scale was adopted, with time measured from the initial diagnosis. Baseline characteristics described above were incorporated to estimate transition-specific hazards, modeled using a Weibull distribution. Under the simplifying assumption that patients with similar clinical characteristics have similar treatment patterns, the model allows estimation of individual-specific transition hazards. We then calculated the cumulative hazard of transition at ten years to predict the transition probabilities by solving the Kolmogorov forward equation through product integral approximation.

Specifically, for each allowed transition from state i to j, the transition intensity was specified with the Weibull hazard function:

$$h_{ij}(t|x) = \lambda_{ij}(x)\,\gamma_{ij}(x)\,t^{\gamma_{ij}(x)-1}$$

where x denotes individual baseline covariates (age, Charlson Comorbidity Index [CCI], and PSA level), and λ_{ij} and γ_{ij} represent the scale and shape parameters, respectively. Covariates were incorporated as predictors influencing both parameters, allowing transition hazards to vary across patient characteristics.

Model evaluation

State transition probabilities were then used to evaluate prediction performance at ten years, with MFS defined as the cumulative probability of remaining in non-metastatic states, and OS defined as the probability of not entering the absorbing death state. Prediction performance was assessed using the C-index on the independent evaluation cohort. For comparison, Cox-PH and RSF models were constructed using the same training cohort and evaluated on the independent 2013 cohort (Table S4). Separate models were fitted for MFS and OS outcomes.

The Cox-PH model estimates the hazard of the event as a function of baseline covariates (age at diagnosis, CCI, and baseline PSA level), under the proportional hazard assumption that the effects of covariates on the hazard remain constant over time. The hazard function is defined as:

$$h(t|Age, PSA, CCI) = h_0(t)\exp\left(\beta_{Age}Age + \beta_{PSA}PSA + \beta_{CCI}CCI\right)$$

where $h_0(t)$ is the unspecified baseline hazard function, and β represents the regression coefficients for the covariates. The model assumes proportional hazards, meaning that covariate effects are constant over time and the hazard ratio between two individuals remains invariant:

$$\frac{h_i(t)}{h_j(t)} = \exp\left(\beta_{Age}\left(Age_i - Age_j\right) + \beta_{PSA}\left(PSA_i - PSA_j\right) + \beta_{CCI}\left(CCI_i - CCI_j\right)\right)$$

Survival probabilities were subsequently obtained from the estimated baseline hazard and covariate effects, with the survival function was derived as:

$$S(t|Age, PSA, CCI) = S_0^{\exp\left(\beta_{Age}Age+\beta_{PSA}PSA+\beta_{CCI}CCI\right)}$$

where $S_0(t)$ denotes the baseline survival function.

RSF models were applied as a non-parametric ensemble approach. Multiple survival trees were constructed using bootstrap sampling of the training data, with each tree grown by recursively partitioning the baseline covariates. Splits were selected to maximize differences in survival outcomes, using the log rank statistic. Within each terminal node, cumulative hazards were estimated using the Nelson–Aalen estimator and converted to survival functions. Individual survival probabilities were obtained by averaging predictions across all trees.

Model performance was compared with a time-inhomogeneous Markov model using C-index and ROC curves, with discrimination quantified by the area under the curve. Longitudinal predictive accuracy was further evaluated by comparing the estimated annual count of event-free people within ten years. Accuracy was defined as one minus the relative error between predicted and observed counts of individuals without progression to metastasis or death (Table S3).

Model application

Based on the predicted probabilities of transfer among health states, we derived the individual time-varying probabilities of progression under three situations: Mild cases, with patients aged $\leq$65 at initial diagnosis, baseline CCI of 0, and PSA levels $\leq$ 4 ng/mL; Severe cases with initial diagnosis aged >80 years, CCI level $\geq$3, and baseline PSA greater than 10 ng/mL; Average risk patients are weighted by the proportion of the combination of baseline age, CCI, and PSA level in the 2010 to 2012 cohort.

QUANTIFICATION AND STATISTICAL ANALYSIS

The multistate Markov models were fit using the *flexsurv* package in R. Parametric survival models for each transition were estimated using the *flexsurvreg* function via maximum likelihood estimation, with optimization performed using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. The estimated transition intensities were then converted into transition probability matrices by numerically solving the Kolmogorov forward equations using the *msfit.flexsurvreg* and probtrans functions. This approach allows estimation of time-dependent state occupancy probabilities and survival outcomes.

The Cox-PH model was fitted using maximum likelihood estimation via the *coxph* function in R. The RSF model was implemented using the *rfsrc* function from the *randomForestSRC* package. The model was trained using 5000 trees, with bootstrap aggregation and log rank splitting used to construct survival trees and generate ensemble-based survival predictions.

Statistical comparison of the covariates between the study and evaluation cohorts was conducted through the Student's t test for the continuous variables and the Chi-square test for the categorical variables. A significance level of 0.05 was adopted for all results interpretation. All statistical analyses were conducted using R software, version 4.2.2 (R Foundation for Statistical Computing, Vienna, Austria), and cross-checked by two independent investigators (Wang J and Jiao Y). The study adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines to ensure transparent and comprehensive reporting.