nature communications

Article

https://doi.org/10.1038/s41467-026-74419-9

A preoperative Artificial Intelligence model to estimate cancer-specific mortality in nonmetastatic kidney cancer patients

Received: 31 January 2025

Accepted: 5 June 2026

Published online: 16 June 2026

Check for updates

Alessandro Larcher[1,7], Alberto Traverso[2,3,7] ✉, Patrick Scuri[2], Umberto Capitanio[1], Giacomo Musso[1], Daniela Canibus[1], Simone Barbieri[2], Luca Caivano[2], Alan Zambello[2], Marco Denti[2], Riccardo Campi[4,5], Sergio Serni[4,5], Salvatore Granata[4], Anna Palmisano[2,6], Davide Vignale[2,6], Fracnesco Montorsi[1,2], Antonio Esposito[2,6], Carlo Tacchetti[2,6,8] & Andrea Salonia[1,2,8]

Surgical resection is the standard treatment for nonmetastatic renal cell carcinoma, yet survival outcomes vary significantly among patients. Current prognostic models lack precision and cannot be applied preoperatively. Here we show the development and validation of a preoperative, interpretable machine learning model to estimate cancer-specific mortality. Using real-world clinical data from 2536 patients and an independent external validation cohort of 580 patients, we combine random survival forests with white-box models to ensure clinical transparency. Our survival tree model relies on exactly eight preoperative features, including tumor size, lymph node involvement, and performance status. We demonstrate that this model outperforms the established GRANT model, achieving a C-index of 0.88 and a Brier score of 0.02 on the external cohort, with notable accuracy in the first year postsurgery. Finally, we provide this tool as a web-based application to facilitate personalized, preoperative risk stratification.

Renal cell carcinoma (RCC) is the most prevalent kidney malignancy, representing around 3% of all cancers[1]. Many renal masses remain asymptomatic until the late disease stages; in this context, the vast majority of RCCs are detected incidentally by abdominal ultrasound or computed tomography (CT) performed for other medical reasons.

Complete surgical resection remains the standard treatment for localized RCC, with partial nephrectomy recommended over radical nephrectomy, to maximally preserve renal function whenever technically feasible.

Despite largely effective treatments, approximately 30% of patients experience recurrence[2], negatively affecting their prognosis, with 5-year cancer-specific survival rates ranging from 67 to 82%, globally[3], prompting international guidelines to advocate for active surveillance based on recurrence risk. In this context, the optimal candidate selection for perioperative systemic therapy remains a subject of intense debate. While adjuvant therapy is established for high-risk patients, the efficacy of neoadjuvant therapy (NAT) remains under investigation. A critical barrier preventing definitive clinical trials for NAT has been the lack of reliable pre-

[1]Division of Experimental Oncology/Unit of Urology, URI, IRCCS Ospedale San Raffaele, Milan, Italy. [2]S-RACE (San-Raffaele Ai CEnter), Vita-Salute San Raffaele University, Milan, Italy. [3]Department of Radiation Oncology (Maastro Clinic), School for Oncology and Reproduction (GROW), Maastricht University Medical Center, Maastricht, the Netherlands. [4]Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy. [5]Unit of Urological Robotic Surgery and Renal Transplantation, University of Florence, Careggi Hospital, Florence, Italy. [6]Experimental Imaging Center, IRCCS San Raffaele Hospital, Milan, Italy. [7]These authors contributed equally: Alessandro Larcher, Alberto Traverso. [8]These authors jointly supervised this work: Tacchetti Carlo, Salonia Aandrea. ✉e-mail: traverso.alberto@hsr.it

surgery models capable of accurately identifying high-risk candidates before the intervention. Current candidate selection often relies on imprecise clinical T-staging or imaging findings alone. As noted by Dibajnia et al.[4], this 'stratification gap' is a primary reason why 65–75% of patients in control arms of adjuvant studies remain disease-free, thereby diluting statistical power. Consequently, there is an urgent unmet need for precise stratification tools capable of identifying high-risk patients in a strictly pre-surgery setting, as highlighted by recent position papers from the Neoadjuvant Kidney Cancer Consortium[5]. Preliminary work on the use of pre-operative prognostic tools has been performed, but using a limited number of pre-operative variables, mainly based on evaluation at imaging of the primary characteristics of the tumor, such as clinical T staging. As an example, Carlo et al.[6] successfully utilized a pre-operative nomogram (Mano et al.[7]) to successfully recruit patients with a > 20% probability of developing metastases. This model was a fine-tuned Cox Proportional Hazard version of the 2008 Leibovich pre-operative nomogram, leaving the possibility of improvements due to the inclusion of a larger number of other potential covariates[8–11].

Thus, research into RCC risk prognostic models aims to personalize treatment. Historically, the tumor, node involvement, and metastasis (TNM) staging system has guided risk assessment; however, recent clinical guidelines emphasize the need for more precise models beyond TNM and tumor grading[12]. Notably, existing models focus on postoperative data, with only a limited number of preoperative risk models, which are urgently needed to advance the personalization of management at an early stage. We reviewed the PubMed literature searching specifically for preoperative models, and we only found two studies that met our search criteria. These studies were limited to variables such as tumor size evaluated at imaging and the presence of symptoms, with a lack of preoperative risk models[13], which are urgently needed to advance the personalization of management at an early stage.

To this aim, the use of Real-World Data (RWD) may offer valuable insights when integrated with responsible Artificial Intelligence (AI) and machine learning (ML) approaches to identify correlations with clinical outcomes and to address clinical decisions for patients with RCC[14].

Responsible AI in healthcare requires avoiding false discoveries, ensuring interpretability, and aligning models with clinical needs. Transparent data selection, robust preprocessing, and interpretable ML methods are essential to build trust and enable clinical adoption. A correct balance between automation and expert oversight is crucial for effectively translating AI innovations into the management of patients with RCC. Finally, ~~novel~~ AI models must be benchmarked against established prognostic models and should be developed into user-friendly clinical applications to facilitate adoption in healthcare settings[15]. Here we aim to develop, externally validate, and deploy a pre-operative machine learning model using real-world data (RWD) to predict mortality risk in RCC patients. Alongside this primary objective, we will benchmark our tool against existing risk prediction models and utilize AI explainability techniques to ensure that the complex algorithms parsing the RWD yield outputs that are clinically interpretable and actionable for real-life practice.

Results

Participants, descriptive data, outcome data

The initial cohort consisted of 3796 patients. After applying exclusion criteria, the final cohort was reduced to 3081 patients. The dataset cardinality before and after preprocessing was as follows:

- Raw DBURI dataset:
  - Input: 3081 patients, 1520 covariates (plus targets).
  - Output: 2536 patients, 206 covariates (plus targets).
- Clinician's dataset:
  - Input: 3081 patients, 126 covariates (plus targets).
  - Output: 2536 patients, 69 covariates (plus targets).

The targets refer to follow-up-related data, including death event indicator (Boolean), time to last follow-up (in months, float), death by cancer (Boolean), and death by other causes (Boolean, always False).

The output data from preprocessing was used to generate Table 1 and the input for the modeling process (Fig. 1). Table 2 shows the cohort description for the external validation. The KM curve for KCSS (Kidney Cancer Specific Survival) of the entire population with curves for prognostic factors are displayed in Supplementary Figs. 2–10. The median follow-up duration was 67 months.

Main results

Raw DBURI and Clinician's dataset trained agnostic exploratory (black box) models show similar performances.

We compared the performance of exploratory models trained on AI-processed RWD with that of models trained on manually curated data. Figure 2 displays metrics for models trained on the Raw DBURI dataset and the Clinician's dataset across 100 MCCV (Monte Carlo Cross Validation) simulations. Mean values were used to compare performance, with standard deviations indicating variability. Higher mean values signify better performance, except for Brier's score, where lower values are preferred.

Based on the overall performance, the metrics obtained on the two datasets are comparable, rejecting the null hypothesis of statistically significant differences ($P < 0{\cdot}05$).

As the performance of exploratory models trained on AI-processed RWD was equivalent to that of models trained on manually curated data, subsequent analysis focused exclusively on models trained on the Raw DBURI dataset.

Extra Survival Trees is the best-performing explainable (white box) pre-operative model.

Internal validation highlights Extra Survival Trees as the best-performing pre-operative model, based on our previously defined empirical formula. Overall, six of the eight T0 models outperformed the T1 model on the C-index (Fig. 3A, B), but all scored worse on the integrated Brier score (Fig. 3C) across all 100 MCCV simulations. Feature counts ranged from 4 (GRANT) to 16. Following the process in Fig. 1, the Extra Survival Trees and the GRANT baseline were retrained on all available internal data for finalization before external validation.

The final pre-operative model included eight prognostic factors: tumor size, lymph node involvement, platelet count, hemoglobin, glomerular filtration rate, age, BMI, and performance status. The pre-operative model explainability (white box approach) highlighted tumor size (measured at imaging based on RECIST criteria), performance status, lymph node metastases (evaluated at imaging), hemoglobin values, and platelet count, as the most relevant features (Fig. 4A).

SHAP analysis revealed trends such as increased mortality risk associated with anemia (low hemoglobin) and thrombocytosis (high platelet count) (Fig. 4, panel A bottom). Table 2 Table 3 highlights the features and coefficients for the GRANT Baseline Cox model, where the relative risk column reflects the hazard ratio (HR) for a one-unit change in continuous variables or the HR for dichotomous variables. Lymph node metastases emerged as the strongest risk factor, followed by Fuhrman grade.

The pre-operative model benchmarked the GRANT baseline model and showed superior performance on external validation.

Using the original external dataset, initially including 720 patients, with 580 patients used for validation after preprocessing, we compared the pre-operative model vs the GRANT baseline. The pre-operative model showed a higher C-index (IPCW: mean 0·87, 95% CI [0·78, 0·95] vs. 0·77, 95% CI [0·64, 0·90]) and cumulative AUC (0·89 vs. 0·73). Both models showed comparable Brier's score IBS (0·017). The

https://doi.org/10.1038/s41467-026-74419-9

Table 1 | Case-mix study for the training cohort

Variable name	Alive		Dead	P-value
N	2407		129	
Anamnesis age (years)	61(52;70)		65(54;71)	**0.018**
Anamnesis CCI age (years)	3(2;5)		4(2;5)	0.039
Anamnesis Charlson comorbidity index	0(0;1)		1(0;2)	0.242
Anamnesis BMI (kg/m^2)	25.6(23.4;28.1)		24.6(22.6;27.5)	0.039
Anamnesis eGFR (ml/min per 1.73 m2)	83.2(69.9;97.4)		77.3(60.6;91.5)	**0.001**
Anamnesis hemoglobin (g/dl)	13.8(12.5;14.9)		12(10.6;13.7)	**<.0001**
Anamnesis creatinine (mg/dl)	0.9(0.8;1.1)		1(0.9;1.3)	**<.0001**
Anamnesis platelets (10^9/l)	226(189;269)		278.5(216;355)	**<.0001**
Anamnesis primary tumor dimension (cm)	4.5(3;6.5)		8(6;11)	**<.0001**
Post-operative Mayo grading	0(0;0)		1(0;1)	**<.0001**
Post-operative tumor dimension (cm)	4.1(2.8;6)		8.6(6;11)	**<.0001**
Post-operative Mayo score	2(0;4)		6(4;8)	**<.0001**
Time to last follow up (months)	60(39;60)		22(10;36)	**<.0001**
	Categories			
Anamnesis ASA score	1	527(22.2%)	18(14%)	**<.0001**
	2	1336(56.3%)	62(48.1%)	.
	3	486(20.5%)	43(33.3%)	.
	4	25(1.1%)	6(4.7%)	.
	5	1(0%)	0(0%)	.
Anamnesis clinical T stage (2009)	T1a	1065(45.9%)	11(8.6%)	**<0.001**
	T1b	723(31.2%)	31(24.2%)	.
	T2a	230(9.9%)	27(21.1%)	.
	T2b	83(3.6%)	14(10.9%)	.
	T3a	170(7.3%)	21(16.4%)	.
	T3b	30(1.3%)	15(11.7%)	.
	T3c	14(0.6%)	9(7%)	.
	T4	3(0.1%)	0(0%)	.
Anamnesis clinical stage (2009)	I	1788(77.1%)	42(32.8%)	**<0.001**
	II	313(13.5%)	41(32%)	.
	III	214(9.2%)	45(35.2%)	.
	IV	3(0.1%)	0(0%)	.
Anamnesis lymphadenopathy positive	FALSE	2191(91%)	88(68.2%)	**<0.001**
	TRUE	216(9%)	41(31.8%)	.
Anamnesis ECOG perform status	0	1036(43.4%)	21(16.4%)	**<0.001**
	1	1110(46.5%)	76(59.4%)	
	2	228(9.6%)	28(21.9%)	
	3	13(0.5%)	3(2.3%)	
Post-operative primary tumor grading	1	243(11.8%)	2(1.6%)	**<0.001**
	2	1380(66.9%)	39(30.5%)	
	3	393(19%)	62(48.4%)	
	4	48(2.3%)	25(19.5%)	

Continuous variables are indicated as median (Q1, Q3), categorical variables as counts (percentage). Statistically significant differences ($P < 0.05$) are highlighted in bold for two-sided Mann-Whitney tests.

pre-operative model demonstrated strong performance, particularly within the first year after surgery, as shown in the AUC plots over time (Fig. 4, panel C). Metrics for both models are summarized in Table 3 Table 4.

The optimal thresholds for risk stratification are displayed on Fig. 4C. The KM curves of the patients stratified are displayed in Fig. 4C. Log-rank tests showed statistically significant split between favorable and intermediate groups ($P < 0{\cdot}005$) and between favorable and unfavorable group ($P < 0{\cdot}005$), while the split between intermediate and unfavorable groups was not significant, likely due to the low number of events in the external dataset.

The preoperative model was deployed as a web-based tool for clinical use, requiring no software installation or expertise by leveraging the S-RACE functionalities. Users can input patient data manually (Fig. 5A). The app provides patient risk group classification (Fig. 5A), KM survival curves compared to the training cohort (Fig. 5A), SHAP individual explanations (Fig. 5B), and variable distributions compared to the training data (Fig. 5C). Further details on app development and usage are provided in the Supplementary Material.

Discussion

We accomplished the primary aim of our study, developing a Web-based preoperative ML risk prediction model for RCC using RWD, to address the still unmet clinical need for a pre-operative ML-based risk prediction model for mortality in RCC.

Our model was trained on a San Raffaele Hospital cohort of 2536 patients and proficiently stratified patients into favorable, intermediate, and unfavorable risk categories. The model was further externally validated on a cohort of 580 patients from another European institution (Careggi Hospital, Florence) as a TRIPOD type 3 model (development of a prediction model using one data set and an evaluation of its performance on separate data)[24]. The web-based solution facilitates the integration of the model into clinical practice while enabling collaborative efforts with other centers to expedite further external validations (TRIPOD type 4 model).

Our study's initial step involved validating our automated data processing pipeline. We did this by comparing the performance of models trained on two datasets from our institutional database (sharing identical inclusion criteria and patient cohort): the Clinician's Dataset, which is the meticulously manually curated benchmark (the gold standard), and the "Raw DBURI Dataset," which was processed comprehensively and automatically by our machine learning pipeline.

The manually curated Clinician's Dataset offers high quality but is not scalable for exploring high-dimensional data. Therefore, our goal was to prove that our automated, AI-driven approach could achieve the same prognostic performance as the model trained on the manually curated data. The resulting parity in performance confirmed that our automated pipeline delivers data quality equivalent to manual curation. This validation was crucial, allowing us to confidently use the complete, larger DBURI dataset for all subsequent analyses and enabling a truly data-driven approach to prognostic modeling. Finally, as a final proof of this robust pipeline, we highlight that out of the eight prognostic features, six were also present in the "Clinician's dataset", but the automated AI pipeline discovered two new features (platelets, hemoglobin) which were not present in the Clinician's dataset.

The model showed robust performance with a C-index of 0·88 and a cumulative AUC of 0·89, and robust calibration with an Integrated Brier's score of 0·02. Notably, the high AUC within the first year indicated that preoperative variables could accurately predict one-year post-surgery mortality risk. Therefore, our model, which uses only preoperative covariates, serves as a potentially useful assessment tool to address early clinical intervention decisions, e.g., proposing neoadjuvant therapy for unfavorable risk patients before surgery or considering less invasive surgeries for favorable risk patients.

https://doi.org/10.1038/s41467-026-74419-9

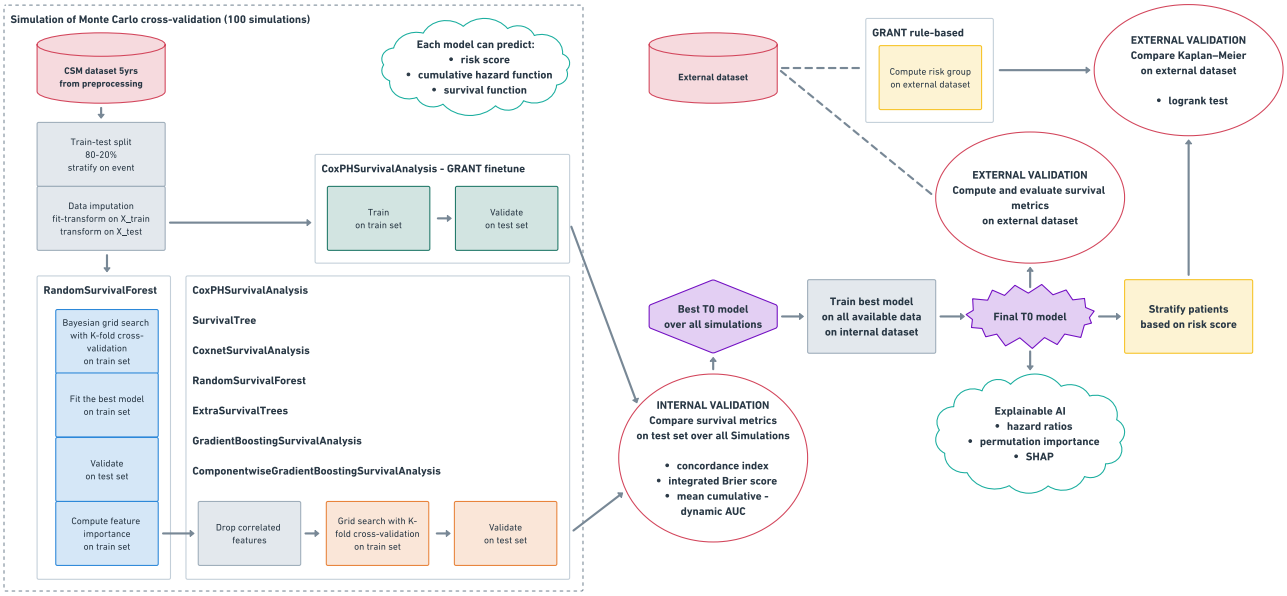

Fig. 1 | Overview of the ML pipeline. First, eight T0 models and the GRANT baseline model are trained on 80% of the internal dataset and validated on the remaining 20%. This procedure is repeated 100 times (dotted panel). Then, the best T0 model is selected, re-trained on the whole dataset, and externally validated on the holdout dataset (Florence dataset), together with the GRANT baseline model.

Despite the complexity of ML algorithms, our preoperative model inference requires only eight easily obtainable clinical features: tumor size, lymph node involvement, platelet count, hemoglobin, glomerular filtration rate, age, BMI, and ECOG performance status. Tumor size, ECOG performance status, and lymph node involvement were found to be the most relevant covariates. Lower hemoglobin and higher platelet counts were linked to poor outcomes, thus confirming previous findings from Xiao et al.[25], Haemmerle et al.[26], and Larcher et al.[27].

To ensure a robust evaluation, we conducted a systematic appraisal of existing prognostic models, including those identified in recent reviews. Our selection of benchmark models (GRANT and SSIGN) was driven by strict inclusion criteria: 1) the availability of reproducible algorithms, 2) alignment with the non-metastatic patient population, 3) the prediction of cancer-specific survival as an endpoint, and 4) endorsement by the European Society of Urology. A granular analysis of 13 prominent models as presented revealed significant barriers to broader benchmarking (see Supplementary Table 6). Specifically, while five identified models predicted KCSS, none functioned as strictly pre-operative tools, as they universally required post-operative pathological data (e.g., necrosis, grade). Furthermore, ~~newer~~ deep-learning approaches (e.g., DeepSurv) lacked publicly available code, precluding reproducibility.

Consequently, we benchmarked our pre-operative model against the SSIGN score (a widely endorsed post-operative model) and the GRANT model. Despite lacking histological inputs, our model outperformed both comparators in external validation (Fig. 4B), confirming its strong discriminative ability relative to established standards.

This comparison shouldn't be framed as a direct competition but as an assessment of two complementary tools designed for different clinical timepoints and purposes. In external validation, our model demonstrated robust and consistent performance, showing improvements in both the C-index and Brier score. In contrast, the GRANT model exhibited a notable discrepancy in its performance metrics—a decline in its C-index (from 0.82 to 0.77) alongside an improvement in its Brier score. This suggests that while our model's predictive accuracy was stable across different evaluation criteria, the GRANT model's performance was sensitive to the shorter follow-up duration of the validation cohort, which favors short-term prediction accuracy over long-term discrimination.

Our pre-operative (T0) model is designed to inform initial risk stratification and surgical planning and/or neoadjuvant therapy before post-surgical pathology is available. In contrast, post-operative (T1) models like GRANT leverage definitive pathological data to refine prognosis and guide subsequent management, such as adjuvant therapy or surveillance intensity. Crucially, our analysis showed that the pre-operative model achieved high discriminative ability, with performance metrics highly competitive with those of the robust post-operative model. This result doesn't imply superiority; rather, it validates the feasibility of making a robust and accurate risk assessment at an early stage using only pre-operative variables. The model's strong performance addresses the unmet clinical need for an effective, early prognostic tool. In a clinical workflow, these models would be used sequentially. Our model provides the initial risk assessment to guide pre-surgical management, which is then updated and refined by a post-operative model once the final pathology results are known. This creates a dynamic, time-dependent approach to patient care, allowing for a more tailored treatment strategy that evolves as definitive information is gathered. In addition, we performed a thorough literature review and found only a few preoperative published models, with ours representing the largest cohort-based study to date[25].

Previously published pre-operative models were developed using limited pre-operative variables, primarily related to tumor size or comorbidities identified during anamnesis. These models relied on manually curated retrospective data from patients treated at local institutions, highlighting the general barriers to collecting large-scale Real-World Data (RWD) for prognostic modeling. In addition, pre-operative research has heavily focused on tasks such as non-invasively predicting the primary tumor's histology or the extent of metastatic disease, enabling disease staging directly from clinical variables[27,28].

Notably, a recent study by Margue et al. introduced UroPredict, a model using both pre- and postoperative covariates based on RWD, which, unlike ours, cannot be applied preoperatively[23].

Our third aim was to evaluate the application of AI explainability techniques to ensure that complex algorithms for parsing RWD produce clinically interpretable and actionable outputs. While ML has the potential to transform RWD analysis through speed and scalability, the

https://doi.org/10.1038/s41467-026-74419-9

Table 2 | Case mix study for the external validation cohort

Variable name	Alive		Dead	P-value
N	566		14	
Anamnesis age (years)	66(57;74)		78.5(74;82)	**<0.001**
Anamnesis CCI age (years)	5(4;7)		7.5(6;9)	**<0.001**
Anamnesis Charlson comorbidity index	3(2;4)		4(3;5)	**0.007**
Anamnesis BMI (kg/m^2)	26.1(23.8;29.4)		24.1(20.9;28.7)	**0.080**
Anamnesis eGFR (ml/min per 1.73 m2)	79.8(65;94.5)		49.8(43.2;57.5)	**<0.001**
Anamnesis hemoglobin (g/dl)	14.2(12.9;15)		12.3(10.8;13.4)	**0.001**
Anamnesis creatinine (mg/dl)	0.9(0.8;1.1)		1.2(1.1;1.6)	**0.001**
Anamnesis platelets (10^9/l)	233(192;273)		239.5(196;278)	0.794
Anamnesis primary tumor dimension (cm)	3.5(2.6;5.2)		6(4.5;9)	**0.001**
Post-operative Mayo grading	2(2;3)		3(2;3)	0.085
Post-operative tumor dimension (cm)	3.4(2.4;4.8)		6.2(4.5;8.5)	**0.003**
Post-operative Mayo score	1(0;3)		4(2;5)	**0.001**
Time to last follow up (months)	34(16;57)		24(5;35)	**0.035**
	Categories			
Anamnesis ASA score	1	119(21%)	1(7.1%)	**0.003**
	2	358(63.3%)	6(42.9%)	
	3	89(15.7%)	7(50%)	
Anamnesis clinical T stage (2009)	T1a	341(60.2%)	2(14.3%)	**<0.001**
	T1b	132(23.3%)	5(35.7%)	
	T2a	42(7.4%)	2(14.3%)	
	T2b	8(1.4%)	2(14.3%)	
	T3a	30(5.3%)	1(7.1%)	
	T3b	4(0.7%)	1(7.1%)	
	T3c	1(0.2%)	0(0%)	
	T4	8(1.4%)	1(7.1%)	
Anamnesis clinical stage (2009)	I	473(83.6%)	7(50%)	**0.007**
	II	50(8.8%)	4(28.6%)	
	III	35(6.2%)	2(14.3%)	
	IV	8(1.4%)	1(7.1%)	
Anamnesis lymphadenopathy positive	FALSE	547(96.6%)	12(85.7%)	0.087
	TRUE	19(3.4%)	2(14.3%)	
Anamnesis ECOG perform status	0	458(80.9%)	6(42.9%)	**0.005**
	1	87(15.4%)	6(42.9%)	
	2	18(3.2%)	2(14.3%)	
	3	3(0.5%)	0(0%)	
Post-operative primary tumor grading	1	102(18.1%)	0(0%)	**0.046**
	2	242(42.8%)	5(35.7%)	
	3	143(25.3%)	8(57.1%)	
	4	78(13.8%)	1(7.1%)	

Continuous variables are indicated as median (Q1, Q3), categorical variables as counts (percentage). Statistically significant differences ($P < 0.05$) are highlighted in bold for two-sided Mann-Whitney tests.

quality of routinely collected clinical data remains a critical challenge. Despite these limitations, RWD—unbiased by manual selection—captures a wide range of covariates, offering opportunities for insights beyond what curated datasets can provide. The ML analysis of RWD data can be performed utilizing a black box or white box modeling approach[29].

Black-box models are complex and difficult to interpret. While they often achieve high accuracy, their lack of transparency can be a significant limitation in healthcare as they raise concerns about trust, ethics, and compliance. On the other hand, white-box models are transparent and easily interpretable and provide clear insights into the decision-making process, making them suitable for healthcare applications where understanding the reasoning behind a diagnosis or treatment recommendation is essential.

However, the dichotomy between white-box and black-box algorithms in healthcare oversimplifies the spectrum of interpretability of AI[29]. Balancing the trade-offs offered by the two approaches is a major challenge for the clinical translation of AI. In our study, we employed a hybrid strategy that leverages the complementary strengths of both approaches. We used a black-box model (Random Survival Forest) to identify hidden signals and patterns within RWD and manually curated datasets, ensuring robust data parsing. Our findings (Fig. 2) demonstrated that automated but rigorous preprocessing through AI enables RWD to achieve performance comparable to that of expertly curated datasets. Based on these results, we next focused on building white-box models using the real-world data (RWD) automatically parsed by the AI.

By leveraging the insights and patterns identified by the black-box model, we developed interpretable models that maintained transparency while retaining clinical relevance. By applying explainability techniques such as feature importance and SHAP values, even advanced algorithms can yield results that are both transparent and clinically meaningful, advancing the integration of real-world evidence into clinical research.

A frequent critique of pre-operative modeling is the absence of definitive histology. However, this constraint must be viewed within the context of a rapidly maturing field focused on non-invasive diagnosis. The reliance on invasive pre-operative biopsy is increasingly being challenged by two parallel technological advancements: AI-driven radiomics and molecular imaging. First, the application of AI to standard-of-care imaging (CT/MRI) has shown significant promise in 'virtual biopsy' tasks. A systematic review and meta-analysis by Mühlbauer et al[30]. reported a significant Log Odds Ratio (OR) of 3.17 for discriminating renal tumor dignity using radiomics, concluding that the application is promising for discrimination of renal tumor dignity. Beyond general discrimination, specific deep learning models have demonstrated high efficacy in granular tasks. For instance, recent validation studies have achieved accuracies as high as 0.95 for differentiating clear cell RCC from oncocytoma (Baghdadi et al.) and 0.90 for four-class histological subtyping (Kilicarlsan et al.) (see Supplementary Tables 4 and 5). Second, complementary to AI, molecular imaging is emerging as a robust diagnostic tool. A seminal 2024 Lancet Oncology study by Shuch et al[31] on clear-cell RCC identified molecular imaging as a highly accurate, non-invasive imaging modality. The authors concluded that this technology has the potential to be 'practice-changing' by allowing for 'earlier diagnosis of clear-cell RCC without unnecessary and invasive procedures. Our pre-operative risk model is designed to function synergistically with these advancements. As non-invasive diagnostics improve their ability to determine histology, the input quality for pre-operative risk models will inherently improve, further solidifying the utility of pre-surgical stratification tools.

Despite numerous strengths, our study is not devoid of limitations. First, it included retrospective data from a single large database, which may introduce a selection bias, though minimized by using

https://doi.org/10.1038/s41467-026-74419-9

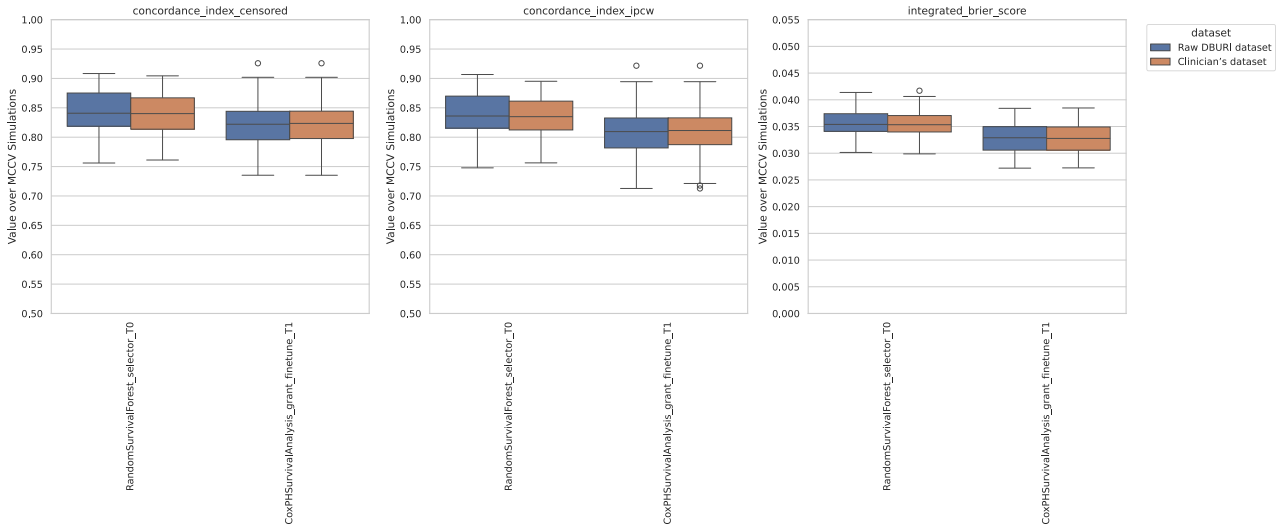

Fig. 2 | Comparison of the results of the Random Survival Forest feature extractors and the GRANT baseline trained end-to-end on both the Raw DBURI dataset (number of patients = 2536) and the Clinician's datasets (number of patients = 2536). The results highlighted no statistically significant differences for all the considered metrics. Box plots show the median (center line), the interquartile range from the 25 to the 75th percentile (box hinges), and whiskers extending to the most extreme data points within 1.5 × IQR beyond the box hinges. Individual data points beyond the whiskers are shown as separate markers.

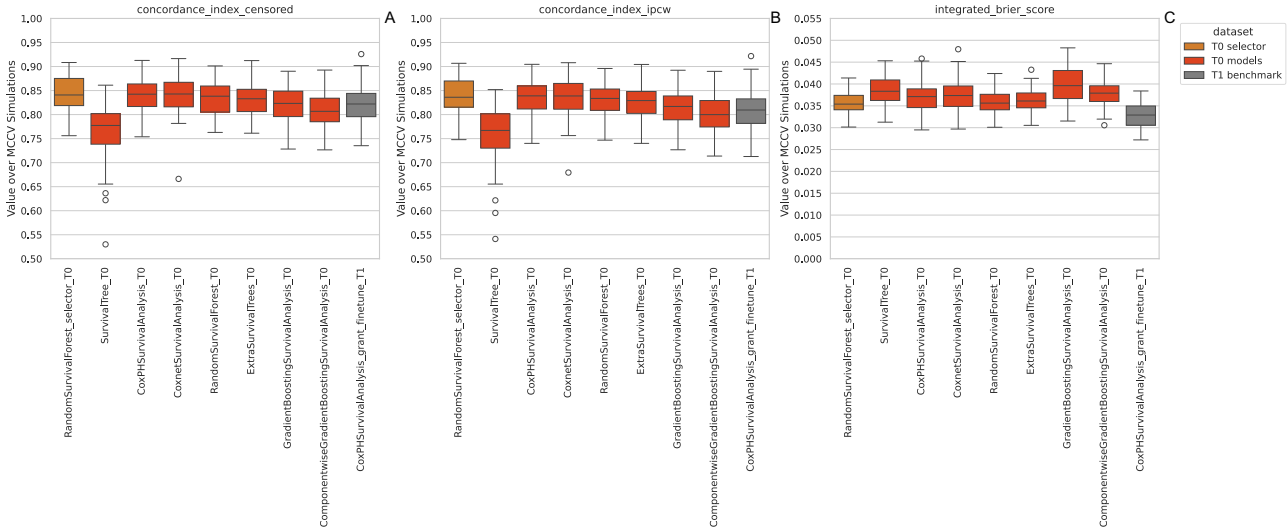

Fig. 3 | Results of internal validation on the Raw DBURI dataset. Results of the trained models (eight T0 models plus the GRANT baseline model in gray) on the internal DB Uri dataset (number of patients = 2536) using the 100 MCCV iterations for the different metrics (C-index censored-top **A**, C-index IPCW-top panel **B**, and Brier's score **C**). Box plots show the median (center line), the interquartile range from the 25 to the 75th percentile (box hinges), and whiskers extending to the most extreme data points within 1.5 × IQR beyond the box hinges. Individual data points beyond the whiskers are shown as separate markers.

routine patient data. Therefore, although the validation on an external matched cohort can mitigate this limitation, a validation on a prospective cohort may further reduce the risk of selection bias.

Methods

The research complies with all relevant regulations, and the study was approved by our local IRB (Protocol N.2007 RENE and subsequent amendments. Written informed consent was not needed for this study due to its investigational retrospective design. The criteria used by the local IRB for informed consent were: Minimal Risk, No Adverse Effect, Infeasibility / Impracticability, and post-participation disclosure. Sex of the participants was included as a potential prognostic factor in the analysis, but no gender analysis was performed since we relied on the biological sex of the participants. Sex was therefore considered to be assigned at birth.

Study design, settings, and participants

Patients with a confirmed diagnosis of RCC who underwent surgery from 1987 to 2020 and were retrospectively and prospectively accrued within our institutional dataset were retrieved from our systems. Surgery types were either partial nephrectomy or radical nephrectomy.

Exclusion criteria were the presence of synchronous metastasis evaluated at imaging (cM +) at the diagnosis of the primary tumor, onset of metachronous cancer, bilateral synchronous RCCs, concomitant hereditary conditions such as Von Hippel–Lindau disease, patients who received neoadjuvant therapy, patients being operated on with total cold ischemia, and the lack of postoperative follow-up data. Patients were evaluated at three time points, as follows: T0 (before surgery) at presentation; T1 (after hospital discharge post remission) for early surgical outcomes; and T2 at 5 years to assess long-term survival.

https://doi.org/10.1038/s41467-026-74419-9

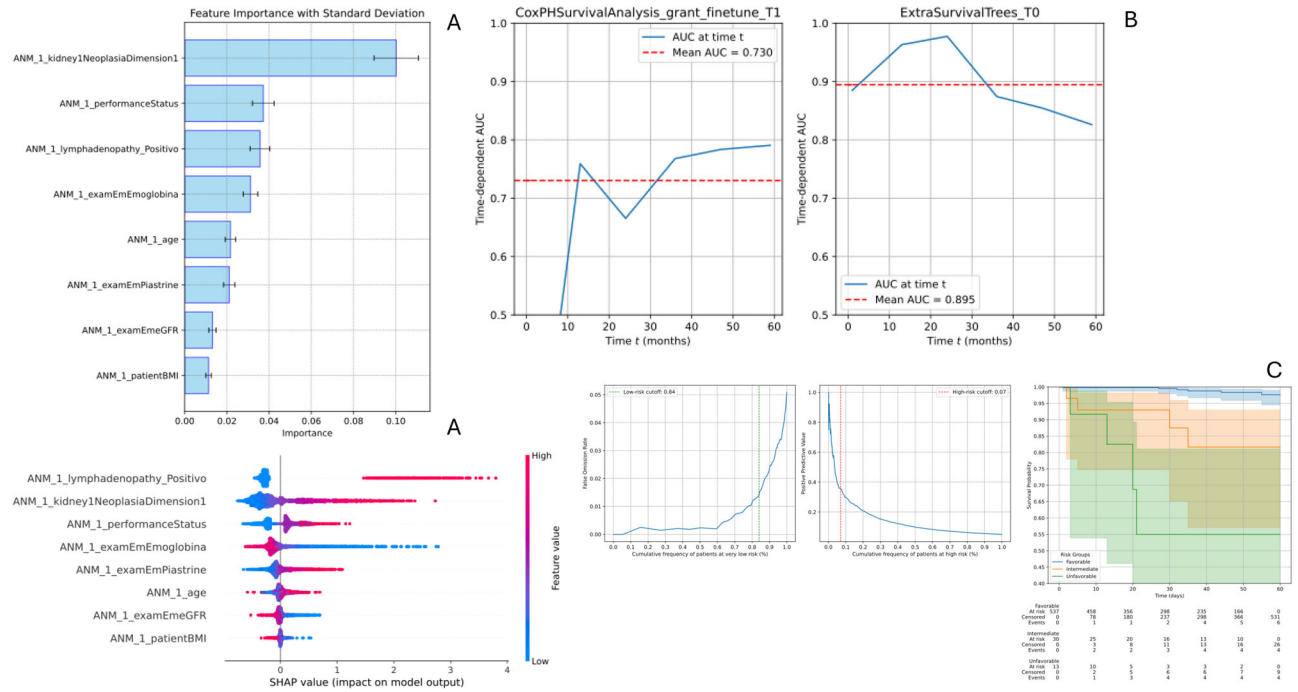

Fig. 4 | T0 model interpretability, external validation performance, and risk group stratification compared to the GRANT baseline T0 model. Feature importance (**A**, top) and SHAP values (**A**, bottom) on the internal dataset (number of patients = 2536). Red color indicates higher values, blue color lower values of the variables. The features are ordered by their importance. Performances of the final models on the external validation dataset: GRANT baseline model (**B** left) and T0 model (**B** right). Stratification into risk groups: cumulative distributions used to identify the optimal thresholds (**C**, left). Risk groups calculated on the external dataset (**C**, right).

Table 3 | Overview of the "GRANT baseline" model trained on the internal dataset, which is a Cox model fined tuned on the training dataset

Variable Name	Variable Description	Coefficient (β)	Relative Risk (exp β)
ANM_1_age_binary	Age True: age greater than 60 years old	0.393	1.481
IST_1_KidneyPathologicalStage2009_binary	Pathological stage True: 1-2-3a False 3b-3c-4	1.280	3.597
IST_1_kidney1PN2009_1_0	Pathological lymph node involvement True: 1-2 False: 0-X	1.710	5.527
IST_1_kidney1Grading_binary	Fuhrman grade True: 3-4v False 1-2	1.609	5

Table 4 | Overview of the performance metrics and calibration of the "GRANT baseline" model and our pre-operative model validated on the external dataset

Metric	Pre-operative model	GRANT baseline	P-value
C-index IPCW	0.87 [0.78-0.95]	0.77 [0.64,0.90]	<0.001
C-index censored	0.88 [0.81-0.96]	0.76 [0.61,0.90]	<0.001
Integrated Brier's score	0.016 [0.008-0.024]	0.017 [0.009-0.026]	<0.001
Cumulative AUC	0.89	0.73	<0.001

CIs are calculated using a 100-sample bootstrapping technique. The comparisons were done using a one-sided Wilcoxon Signed-Rank Test.

Patients were followed up until death, and the causes of death were assessed for all patients by consulting the patients' medical records. Five-year kidney cancer-specific survival (KCSS) after surgery was the main endpoint. Likewise, the same inclusion criteria were used for an external validation dataset of RCC patients treated and accrued at another tertiary-care European Urological Department (Careggi Hospital, Florence).

Statistics and reproducibility

No statistical method was used to predetermine sample size. No data were excluded from the analyses. The experiments were not randomized.

Quantitative variables (data pre-processing) and statistical methods

The dataset was sourced from the Urological Research Institute—San Raffaele Hospital, referred to as the URI dataset. Two versions were utilized: the Raw DBURI dataset (unprocessed) and the Clinician's dataset, which underwent initial feature selection, engineering, and refinement by a urologist with extensive experience in epidemiology and outcome research, who selected covariates based on evidence from the literature and clinical guidelines. Both datasets are structured as feature tables and follow matched inclusion criteria. The

https://doi.org/10.1038/s41467-026-74419-9

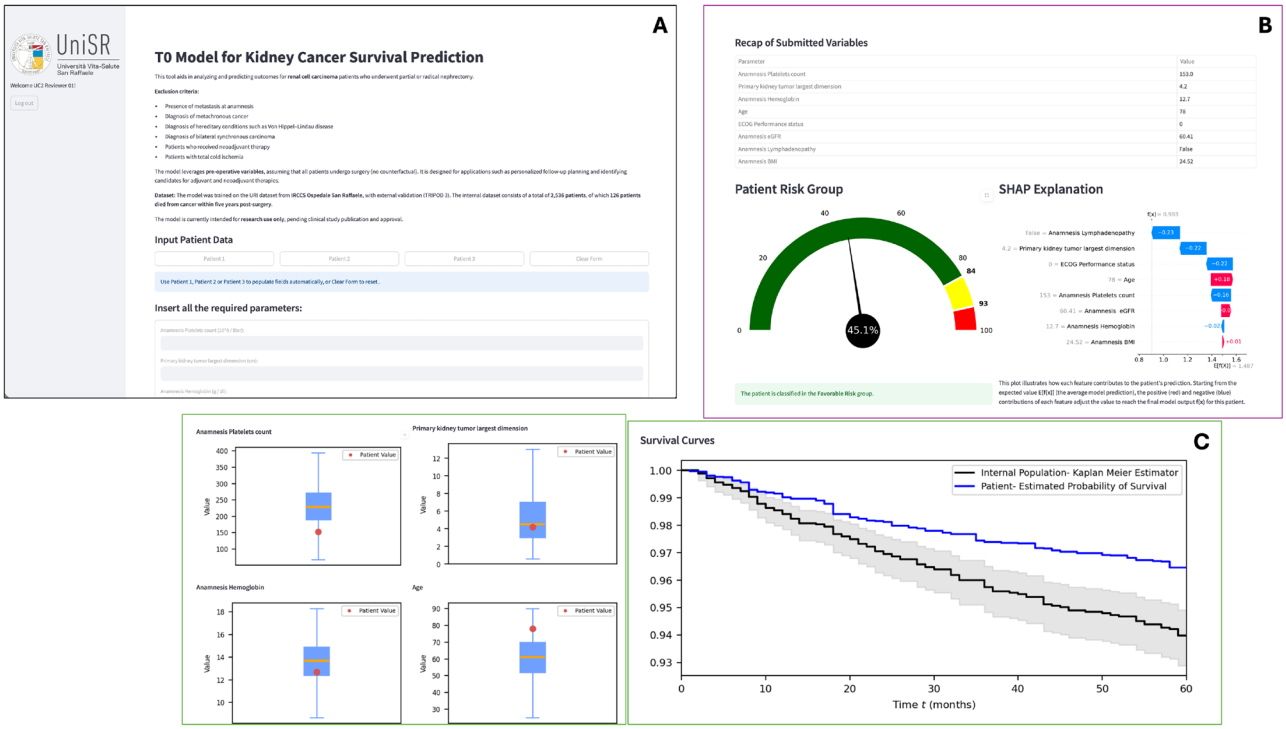

Fig. 5 | Deployment of the model as a clinical interface. Overview of the Web-based calculator (**A**). Data can be inserted manually using the insertion form (**A**, bottom). For each individual patients, results are reported: risk assignment and individual KM curve compared to the KM of the cohort used to train the model (panel A, bottom), SHAP individual explanations (**B**), and distribution of single patient's values with respect to the training cohort (**C**) (number of patients = 2536). Web application custom-developed in Python 3.12.9 with the libraries Streamlit 1.47.0, Matplotlib 3.8.2, SHAP 0.46.0, and deployed in Azure cloud. Box plots show the median (center line), the interquartile range from the 25th to the 75th percentile (box hinges), and whiskers extending to the most extreme data points within 1.5 × IQR beyond the box hinges. Individual data points beyond the whiskers are shown as separate markers. Figures generated using Python 3.12.9 with libraries Matplotlib 3.8.2 and sciktlearn. The code use to generate the figure is downloadable at the Zenodo link provided in the software availability statement.

Raw DBURI dataset includes thousands of variables, spanning the following domains at two time points (T0-before surgery and T1-post surgery): demographics, comorbidities, tumor information, and surgery records. Details on the data are shown in Supplementary Table 1.

The Clinician's dataset manually refines these variables by means of processing raw data with statistical software (e.g. SPSS and R), including established prognostic factors from the literature.

The datasets were loaded in the San Raffaele cloud-based data science platform, called S-RACE (San Raffaele Center of AI). The platform has three main functionalities: i) universal data platform (ingestion); ii) clinician AI Hub (exploration), and iii) data science lab (modeling). S-RACE allows for performing end-to-end clinical data science research projects, taking care of data pseudo-anonymization, data pre-processing and structuring, and development of reproducible ML pipelines for data analyses, thus providing the opportunity to develop AI solutions balancing advanced capabilities with explainability. Details on the S-RACE architecture and the main building blocks are shown in Supplementary Fig. 1.

Following common procedures indicated by the literature[16], both datasets underwent a preprocessing phase to prepare for ML analysis. Details are provided in the Supplementary Table 2. Most preprocessing steps were automated, with manual input required for steps 1 and 3, under the supervision of a clinical expert. A multivariate iterative imputation strategy was employed to fill missing values, using a round-robin approach that models each variable as a function of others (MICE-algorithm)[17]. Prior to imputation, we assessed the missing data mechanism to ensure suitability for MICE. Following the methodology described by Little[18] and Sterne et al.[17], we generated missingness indicators (binary variables where 1 indicates a missing value and 0 indicates an observed value) for all incomplete covariates. We then examined the correlation matrix between these indicators and the observed values of other covariates and the outcome. The analysis revealed no strong systematic correlations that would suggest a Missing Not At Random (MNAR) mechanism, thereby supporting the plausibility of the Missing At Random (MAR) assumption required for valid multiple imputation.

Machine learning pipeline

Figure 1 highlights the study design and analysis pipeline.

Data partitioning. A stratified 80/20 train-test split, using 100 simulations for Monte Carlo cross-validation (MCCV)[19] ensured robust model generalization. Hyperparameter optimization was conducted with 10-fold cross-validation on the training set, effectively creating a train-validation-test split for each MCCV simulation.

Model design. We used models from the Python v.3.10.13 library "scikit-sksurv" with a two-phase ML process: 1) a Random Survival Forest-based feature selection, ranking features by importance; 2) multiple model pipelines (each with MCCV) using the selected features. A Cox proportional hazards model (GRANT features, fine-tuned on our internal dataset) served as a baseline. Hyperparameter optimization (10-fold cross-validation), feature selection, and dimensionality reduction (Spearman correlation <0·6, max 16 features) were performed on both the feature selector and final models. Eight models (were trained via grid search, maximizing the concordance index (C-index). The best model was chosen using a weighted average of normalized C-index, integrated Brier score, and feature count over the 100 MCCV (weights: 0·5, 0·5, 0·25, respectively, with inverse normalization for lower-is-better metrics). This model was then refitted to the full dataset.

Evaluation metrics. The performance of survival models on both the internal dataset (test set) and the external dataset was assessed using four key metrics measuring classification performances (C-index) and calibration (Brier's score) (detailed description in Supplementary Table 3). The main results focus on the C-index with inverse probability of Censoring weights (IPCW), shown to be less prone to optimistic estimations[20]. On the external validation, the C-index and Brier's score are computed with 100 bootstraps, allowing for confidence interval estimation.

Explainability. The final models are interpreted, after internal validation, based on.

I. Intrinsic explainability for Cox models: coefficients and hazard ratios.
II. Permutation importance with 1000 repeats[21]: evaluating the importance of each feature in a ML model by measuring the impact on model performance when a feature's values are randomly shuffled. The logic is that if a feature is important, shuffling its values will degrade the model's performance, as it disrupts the relationship between the feature and the target variable.
III. Global SHAP (SHapley Additive exPlanations)[22] on all the patients: explaining, based on game theory, the output of machine learning models by attributing the contribution of each feature to a model's prediction.

Risk stratification. Patients were stratified to three risk groups for cancer-specific mortality based on the risk probabilities as obtained from our preoperative model. The optimal threshold for stratification was determined by adopting the approach used in the UroPredict study[23]. The thresholds were set to obtain a large proportion of patients with a very low (favorable risk group) and a significant proportion of patients with a high actual relapse rate (unfavorable risk group). These stratification thresholds were determined using the training dataset and then applied to patients in the external dataset.

The KCSS curves for the three risk groups were estimated using the Kaplan-Meier (KM) method and compared using log-rank tests. To determine the threshold for favorable risk patients, we decided to plot the false omission rate (1- Negative Predictive Value) as a function of the cumulative frequency of patients in the very low-risk group by varying the risk threshold. Compared to the above-cited study, we did not select only the threshold, such as there is a significant increase in the false omission rate, but also guaranteeing that the number of included patients was in line with the reported median KCSS for non-metastatic patients (84%)[4]. We could then use a similar strategy with the positive predictive value (PPV) to determine an unfavorable risk group of patients. We look for a PPV plateau to determine the risk thresholds (threshold 0.07).

External validation. The preprocessing steps for the external dataset involved: mapping the categories as in the internal dataset (e.g., TNM staging T1a as 1), dropping missing values for event, dropping unknown causes of death, altering deaths at 0 months to 1 month (1 patient), and filtering for KCSS while censoring events after 5 years. Figure 1 depicts the end-to-end experimental design.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are not publicly available due to the pseudonymized nature of individual-level records, which could compromise research participant privacy and consent. Source data are provided with this paper for figures that do not involve individual patient-level information; figures derived from individual-level data – including Kaplan–Meier survival estimates and SHAP beeswarm plots – cannot be accompanied by source data files. Source data are provided with this paper.

Code availability

The code used to train and validate the model can be accessed at the following GitHub repository https://github.com/AI-UniSR/rcc_risk_prediction_ml under the MIT license and associated with the Zenodo https://zenodo.org/records/19093594 at ref. 32.

References

1. Siegel, R. L., Miller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics. *CA Cancer J. Clin.* **73**, 17–48 (2023).
2. Lucca, I., Klatte, T., Fajkovic, H., de Martino, M. & Shariat, S. F. Gender differences in incidence and outcomes of urothelial and kidney cancer. *Nat. Rev. Urol.* **12**, 585–592 (2015).
3. Larcher, A. et al. Epidemiology of renal cancer: incidence, mortality, survival, genetic predisposition, and risk factors. *Eur. Urol.* **88**, 341–358 (2025).
4. Dibajnia, P., Cardenas, L. M. & Lalani, A. K. A. The emerging landscape of neo/adjuvant immunotherapy in renal cell carcinoma. *Hum. Vaccines Immunotherapeutics* **19**, 2178217 (2023).
5. Bex, A. et al. A call for a neoadjuvant kidney cancer consortium: lessons learned from other cancer types. *Eur. Urol.* **87**, 385–389 (2025).
6. Carlo, M. I. et al. Phase II study of neoadjuvant nivolumab in patients with locally advanced clear cell renal cell carcinoma undergoing nephrectomy. *Eur. Urol.* **81**, 570–573 (2022).
7. Mano, R. et al. Preoperative nomogram predicting 12-year probability of metastatic renal cancer - evaluation in a contemporary cohort. *Urol. Oncol.* **38**, 853.e1–853.e7 (2020).
8. Fallara, G. et al. How to optimize the use of adjuvant pembrolizumab in renal cell carcinoma: which patients benefit the most? *World J. Urol.* **40**, 2667–2673 (2022).
9. Bedke, J. et al. The 2022 updated european association of urology guidelines on the use of adjuvant immune checkpoint inhibitor therapy for renal cell carcinoma. *Eur. Urol.* **83**, 10–14 (2023).
10. Marandino L., et al. Neoadjuvant and adjuvant immune-based approach for renal cell carcinoma: pros, cons, and future directions. *Eur. Urol. Oncol.* **25**, S2588-931100211-6 (2024).
11. Cerrato, C. et al. Effect of sex on the oncological outcomes in response to immunotherapy and antibody-drug conjugates in patients with urothelial and kidney cancer: a systematic review and a network meta-analysis. *Eur. Urol. Oncol.* **7**, 1005–1014 (2024).
12. Escudier, B. et al. Renal cell carcinoma: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **30**, 706–720 (2019).
13. Harrison, H. et al. Risk prediction models for kidney cancer: a systematic review. *Eur. Urol. Focus* **7**, 1380–1390 (2021).
14. Purpura, C. A., Garry, E. M., Honig, N., Case, A. & Rassen, J. A. The role of real-world evidence in FDA-approved new drug and biologics license applications. *Clin. Pharm. Ther.* **111**, 135–144 (2022).
15. Wee, L. et al. Reporting Standards and Critical Appraisal of Prediction Models. In: Kubben P., Dumontier M., Dekker A., curatori. Fundamentals of Clinical Data Science [Internet]. Cham: Springer International Publishing; 2019 [citato 14 ottobre 2024]. p. 135–150. Disponibile su: http://link.springer.com/10.1007/978-3-319-99713-1_10.
16. Van, K. S. M. J., Dankers, F. J. W. M., Traverso, A. & Wee L. Preparing data for predictive modelling. In: Kubben P., Dumontier M., Dekker A., curatori. Fundamentals of Clinical Data Science [Internet]. Cham: Springer International Publishing; 2019 [citato 6 novembre 2024]. p. 75–84. Disponibile su: http://link.springer.com/10.1007/978-3-319-99713-1_6.

17. Azur, M. J., Stuart, E. A., Frangakis, C. & Leaf, P. J. Multiple imputation by chained equations: what is it and how does it work? *Int. J. Methods Psychiatr. Res.* **20**, 40–49 (2011).
18. Little, R. J. A. A test of missing completely at random for multivariate data with missing values. *J. Am. Stat. Assoc.* **83**, 1198–1202 (1988).
19. Shan, G. Monte Carlo cross-validation for a study with binary outcome and limited sample size. *BMC Med Inf. Decis. Mak.* **22**, 270 (2022).
20. Schober, P. & Vetter, T. R. Survival analysis and interpretation of time-to-event data: the tortoise and the hare. *Anesthesia Analgesia* **127**, 792–798 (2018).
21. Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics* **26**, 1340–1347 (2010).
22. Louhichi, M., Nesmaoui, R., Mbarek, M. & Lazaar, M. Shapley values for explaining the black box nature of machine learning model clustering. *Procedia Computer Sci.* **220**, 806–811 (2023).
23. Margue, G. et al. UroPredict: machine learning model on real-world data for prediction of kidney cancer recurrence (UroCCR-120). *npj Precis. Onc.* **8**, 45 (2024).
24. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann. Intern. Med.* **162**, 55–63 (2015).
25. Xiao, R. et al. Preoperative anaemia and thrombocytosis predict adverse prognosis in non-metastatic renal cell carcinoma with tumour thrombus. *BMC Urol.* **21**, 31 (2021).
26. Haemmerle, M., Stone, R. L., Menter, D. G., Afshar-Kharghan, V. & Sood, A. K. The platelet lifeline to cancer: challenges and opportunities. *Cancer Cell* **33**, 965–983 (2018).
27. Larcher, A. et al. When to perform preoperative chest computed tomography for renal cancer staging. *BJU Int.* **120**, 490–496 (2017).
28. Larcher, A. et al. When to perform preoperative bone scintigraphy for kidney cancer staging: indications for preoperative bone scintigraphy. *Urology* **110**, 114–120 (2017).
29. Loyola-Gonzalez, O. Black-box vs. white-box: understanding their advantages and weaknesses from a practical point of view. *IEEE Access* **7**, 154096–154113 (2019).
30. Mühlbauer, J. et al. Radiomics in renal cell carcinoma-A systematic review and meta-analysis. *Cancers (Basel)* **13**, 1348 (2021).
31. Shuch, B. et al. 89Zr]Zr-girentuximab for PET-CT imaging of clear-cell renal cell carcinoma: a prospective, open-label, multicentre, phase 3 trial. *Lancet Oncol.* **25**, 1277–1287 (2024).
32. Larcher, A. et al. A preoperative artificial intelligence model to estimate cancer specific mortality in nonmetastatic kidney cancer patients. *Zenodo* https://doi.org/10.5281/zenodo.19093594 (2026).

Author contributions

L.A., A.T., and S.P.: conceptualization, data collection, analysis, writing, revision, figures, tables. C.D.: data collection, writing, revision. C.U., M.G., B.S., C.R., S.S., G.S., P.A., and V.D.: conceptualization, data collection, writing, revision. C.L., Z.A., and M.D.: software development, writing, revision. M.F., E.A., T.C., and S.A.: conceptualization, supervision, writing, revision.

Funding

This work was supported by the Italian Ministry of Research, under the complementary actions to the NRRP "D34health - Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care" Grant (# PNC0000001) (TC). The funding source had no role for writing or decision to submit this manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-74419-9.

Correspondence and requests for materials should be addressed to Alberto Traverso.

Peer review information *Nature Communications* thanks Bingqing Xie, James Jones and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.