BJU Int 2022 doi:10.1111/bju.15673

Review

BJUI
BJU International

Risk models for recurrence and survival after kidney cancer: a systematic review

Juliet A. Usher-Smith[1], Lanxin Li[2], Lydia Roberts[2], Hannah Harrison[1], Sabrina H. Rossi[3], Stephen J. Sharp[4], Carol Coupland[5], Julia Hippisley-Cox[6], Simon J. Griffin[1], Tobias Klatte[7] and Grant D. Stewart[8]

[1]Primary Care Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, [2]School of Clinical Medicine, University of Cambridge, Cambridge, [3]Department of Oncology, Addenbrooke's Hospital, University of Cambridge, Cambridge, [4]MRC Epidemiology Unit, University of Cambridge, Cambridge, [5]School of Medicine, University of Nottingham, Nottingham, [6]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, [7]Royal Bournemouth Hospital, Bournemouth, and [8]Department of Surgery, Addenbrooke's Hospital, University of Cambridge, Cambridge, UK

J.A.U.-S., L.L. and L.R. were equal contributors to this work.

Objective

To systematically identify and compare the performance of prognostic models providing estimates of survival or recurrence of localized renal cell cancer (RCC) in patients treated with surgery with curative intent.

Materials and Methods

We performed a systematic review (PROSPERO CRD42019162349). We searched Medline, EMBASE and the Cochrane Library from 1 January 2000 to 12 December 2019 to identify studies reporting the performance of one or more prognostic model(s) that predict recurrence-free survival (RFS), cancer-specific survival (CSS) or overall survival (OS) in patients who have undergone surgical resection for localized RCC. For each outcome we summarized the discrimination of each model using the *C*-statistic and performed multivariate random-effects meta-analysis of the logit transformed *C*-statistic to rank the models.

Results

Of a total of 13 549 articles, 57 included data on the performance of 22 models in external populations. *C*-statistics ranged from 0.59 to 0.90. Several risk models were assessed in two or more external populations and had similarly high discriminative performance. For RFS, these were the Sorbellini, Karakiewicz, Leibovich and Kattan models, with the UCLA Integrated Staging System model also having similar performance in European/US populations. All had *C*-statistics ≥0.75 in at least half of the validations. For CSS, they the models with the highest discriminative performance in two or more external validation studies were the Zisman, Stage, Size, Grade and Necrosis (SSIGN), Karakiewicz, Leibovich and Sorbellini models (*C*-statistic ≥0.80 in at least half of the validations), and for OS they were the Leibovich, Karakiewicz, Sorbellini and SSIGN models. For all outcomes, the models based on clinical features at presentation alone (Cindolo and Yaycioglu) had consistently lower discrimination. Estimates of model calibration were only infrequently included but most underestimated survival.

Conclusion

Several models had good discriminative ability, with there being no single 'best' model. The choice from these models for each setting should be informed by both the comparative performance and availability of factors included in the models. All would need recalibration if used to provide absolute survival estimates.

Keywords

recurrence, renal cell cancer, risk prediction, survival, prognosis, #kcsm, #KidneyCancer, #uroonc

Introduction

International guidelines recommend that the surveillance of individuals with localized clear-cell RCC (ccRCC) should be stratified according to the risk of developing recurrence. The AUA [1] and the National Comprehensive Cancer Network (NCCN) [2] recommend stratification based on TNM staging. The European Society for Medical Oncology (ESMO) [3] and European Association of Urology (EAU) [4] provide a strong recommendation for the use of other prognostic models, considering them more accurate than TNM stage or grade alone for predicting clinically relevant outcomes. A large number of such prognostic models have been developed and many have been compared with each other in external validation studies. Existing reviews of these models [5,6], however, are non-systematic and do not provide data on direct comparisons between the models. Both the ESMO and EAU state that there is insufficient evidence to recommend one prognostic model over another, with the ESMO giving the examples of the UCLA Integrated Staging System (UISS) and the Stage, Size, Grade and Necrosis (SSIGN) score, and the EAU citing the UISS, Leibovich and Grade, Age, Nodes and Tumour (GRANT) models as being the current most relevant prognostic models for ccRCC. The decision on which model to use is, therefore, left to the individual clinician, with potential for variation in patient care.

Recent advances in adjuvant treatment for ccRCC, in particular the KEYNOTE-564 trial which showed a significant disease-free survival benefit for pembrolizumab over placebo [7], additionally make it likely that, for the first time, adjuvant immunotherapy will be recommended to patients at high risk of recurrence in the near future. Prognostic models will therefore become even more important as they will be needed to identify high-risk patients likely to benefit from such adjuvant therapy.

To inform future guidelines both for surveillance and adjuvant immunotherapy and to support clinicians to make an informed choice of model, we performed the first systematic comparison of the performance of prognostic models that provide estimates of recurrence or survival after ccRCC treated with surgery with curative intent.

Materials and Methods

We performed this review according to a published protocol (PROSPERO 2019 CRD42019162349 Available from: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42019162349) and in line with guidance for systematic reviews of prediction model performance [8]. The results are reported in accordance with the TRIPOD guidelines [9].

Search Strategy

We systematically searched Medline, EMBASE and the Cochrane Library for studies published from 1 January 2000 to 12 December 2019 using a combination of subject headings incorporating 'kidney cancer/renal cell cancer', 'recurrence/survival/prognosis' and 'prediction/model/score' (Tables S1 and S2). The search was extended by manually screening the reference lists and electronically searching for citations of included papers.

Inclusion Criteria

We included peer-reviewed studies that reported a quantitative measure of the performance of one or more risk model(s) including a combination of ≥2 risk factors to predict at least one of the outcomes of interest at an individual level in patients after surgical resection for localized RCC. The outcomes of interest were drawn from the DATECAN guidelines for time-to-event endpoints in RCC clinical trials [10] and included recurrence-free survival (RFS), cancer-specific survival (CSS) and overall survival (OS). RFS included metastasis-free survival, local recurrence-free survival, progression to metastatic disease and recurrence of disease. To avoid overestimates of performance due to overfitting, we included only studies measuring the performance of models in a population distinct from the model development population (external validation) in the primary analysis. To inform future models and identify potentially promising prognostic markers, we included studies for a secondary analysis that reported the performance of an existing model in an external population alongside the performance of that model plus any additional prognostic markers in the same population.

We excluded studies in which it was not possible to separate patients with localized disease from those with metastatic disease at the time of recruitment and studies including only specific groups, for example, studies including only patients with high-grade or locally advanced disease and those limited to transplant recipients, individuals with inherited renal cancer syndromes, or non-clear-cell subtypes of RCC.

Study Selection

Title and abstract screening were performed using Rayyan (https://rayyan.ai). After piloting the inclusion and exclusion criteria to achieve >98% agreement, titles and abstracts were assessed by one author, with a random 10% checked by a second. Full-text screening was performed by four reviewers in two stages. In the first stage, review articles, conference abstracts, studies with no performance measures, duplicate studies and studies including only single risk factors were excluded. In the second stage, the remaining papers were screened by two reviewers against the other inclusion criteria.

Data Extraction

Data were extracted directly into data tables by two authors. A random 10% of data were additionally checked by a third. For studies that reported the stepwise performance of models, only the model with the best performance was extracted. Where studies included estimates of discrimination for multiple durations, only data for the longest time period were extracted. Where the same risk model was assessed in participants recruited from the same site during the same time period in more than one study, we extracted only the performance data from the study with the greatest number of outcomes.

Risk-of-Bias Assessment

A risk-of-bias (RoB) assessment was performed separately for each external validation, model and outcome using the PROBAST tool (Method S1) [11]. We extracted data relevant to the assessment of RoB at the same time as data extraction. One author then completed the RoB assessment, with a random 10% checked by a second author.

Data Synthesis

Data were synthesized separately for the three outcomes (RFS, CSS and OS). The discrimination for each model was summarized graphically with the *C*-statistic. For each model for each outcome we also estimated heterogeneity in model performance using the I^2 statistic [12] within the 'metan' command in Stata with the logit transformed *C*-statistics [8,13] and restricted maximum likelihood estimation.

As the heterogeneity across the studies was high (up to 95%) we did not estimate pooled *C*-statistics. To enable us to rank the relative discrimination of the models and incorporate both direct and indirect evidence from risk model comparisons across the studies, we performed multivariate random-effects meta-analyses, again using the logit transformed *C*-statistic, using the 'mvmeta' command in Stata [14,15]. For these analyses we used the Riley method to estimate within-study correlations [16] and used the conventional assumption that all the pairwise between-study correlations were 0.5. We present the borrowing of strength, which is the percentage weight in the meta-analysis that is given to the indirect evidence [17], the mean rank, which is the average ranking for each model included in the analysis [18], and the surface under the cumulative ranking curve (SUCRA), which is the mean rank scaled from 0 to 1 to enable comparisons across outcomes, from that analysis. Studies where it was not possible to calculate a confidence interval of the *C*-statistic were excluded from that analysis.

To explore potential sources of heterogeneity among the studies we performed subgroup analyses by study geographical region (Europe/US and Asia) and, where there were eight or more external validations of the same model, we used meta-regression to explore the association between study-level characteristics (event rate, proportion of participants with ccRCC, baseline year of recruitment and duration over which risk was predicted) and the *C*-statistic.

The measures of calibration, estimated survival for patients in different categories of risk defined by the models and increase in performance of risk models with the addition of other prognostic markers are summarized descriptively.

Results

Our search identified 13 549 articles. Of these, 75 met our inclusion criteria (Fig. S1 and Table S3). The most common reasons for articles to be excluded at full-text review were that the cohort included patients with metastatic disease or specific groups of patients, such as only those with low-risk or high-risk disease, or that the study was not an external validation. Fifty-seven included data on the performance of 22 risk models in an external population and 40 included data on the improvement in performance of previously published risk models with the addition of one or more additional prognostic markers. Most studies recruited participants from single centres, with all but two [19,20] recruiting participants retrospectively. The RoB assessments for each study for each external validation are detailed in Tables S4–S6. Of the 150 validations assessed (69 RFS, 38 CSS and 43 OS), 95 were rated as having high RoB, 49 as having unclear RoB (typically due to a lack of clear reporting) and six as having low RoB. Across the four domains assessed (Method S1), issues with analysis were most frequently noted. Common problems included the management of participants lost to follow-up and the use of datasets with very few events (<50).

Details of the risk factors included and scoring for each of the 22 risk models are given in Table 1 [21,22,23,24–30,31–40,41]. The majority included pathological or clinical prognostic factors that are likely to be available in routine clinical practice. Two included genetic risk factors (Wei et al. [26] and the Recurrence score [24]), one included molecular markers [23] and five included biochemical markers (e.g. albumin and C-reactive protein [CRP]) that may be available in some settings (CONtrolling NUTritional status (CONUT) [36], glasgow prognostic score (GPS) [39], modified GPS (mGPS) [30], prognostic nutritional index (PNI) [41] and Chen et al. [21]).

Recurrence-free Survival

Discrimination

A total of 36 studies reported 69 *C*-statistics for external validations of 19 models for RFS after surgery (Table S4). The median duration of follow-up was reported for 59 of the external validations and ranged from 33 to 128 months, with

Table 1 Details of included risk models.

Risk model	Country of development	Development population	Original outcome	Risk factors/prognostic factors included	Risk groups/prognostic groups	Risk factors available
Chen et al. [21]	China	ccRCC	OS	1. T stage 2. Neutrophil to lymphocyte ratio 3. Monocyte to lymphocyte ratio 4. Albumin to globulin ratio	Nomogram giving continuous quantification of risk	Potentially
Cindolo et al. [31]	France, Italy	RCC	RFS	1. Clinical size 2. Clinical presentation (symptomatic vs asymptomatic)	RRF = (1.28 × presentation (asymptomatic = 0; symptomatic = 1) + (0.13 × clinical size)) Good prognosis group: RRF ≤1.2 Poor prognosis group: RRF >1.2	Y
CONUT [36]	Spain	Nutrition risk	Risk of hospital malnutrition	1. Serum albumin 2. Total lymphocytes 3. Cholesterol	Total score calculated between 0 and 12 Normal: 0–1 Light: 2–4 Moderate: 5–8 Severe: 9–12	Potentially
GPS [39]	UK	Inoperable non-small-cell lung cancer	OS	1. CRP 2. Albumin	*Score* Elevated CRP (>10 mg/L) and hypoalbuminaemia (<35 g/L) = 2 Elevated CRP (>10 mg/L) or hypoalbuminaemia (<35 g/L) = 1 CRP ≤10 mg/L and albumin ≥35 g/L = 0	Potentially
GRANT [33]	USA, Canada	RCC	OS, RFS	1. Pathological tumour size 2. Pathological nodal status 3. Fuhrman grade 4. Age	Number of unfavourable risk factors is summed (0–4) Favourable group: score 0–1 Unfavourable group: score ≥2	Y
Jeong et al. [22]	South Korea	ccRCC	RFS	1. Tumour size 2. Macroscopic appearance 3. Age	Score range 0–18 Low risk: score ≤3.5 Intermediate risk: score >3.5 and ≤10.5 High risk: score >10.5	Y
Karakiewicz et al. [34]	France, Italy	RCC	CSS	1. T stage 2. N stage 3. M stage 4. Tumour size 5. Fuhrman grade 6. Symptom classification	Nomogram giving continuous quantification of risk	Y
Kattan et al. [27]	USA	RCC	RFS	1. Pathological tumour stage 2. Tumour size 3. Histology 4. Symptoms	Nomogram giving continuous quantification of risk	Y
Klatte et al. [23]	USA	ccRCC	RFS	1. T classification 2. ECOG PS 3. Ki-67 expression 4. p53 expression 5. Epithelial VEGFR-1 expression 6. Endothelial VEGFR-1 expression 7. Epithelial VEGF-D expression	Three risk groups based on total points assigned by the nomogram Low-risk group: ≤120 points Intermediate-risk group: 121–175 points High-risk group: >175 points	N
Leibovich et al. [38]	USA	ccRCC	RFS	1. Pathological T stage 2. Regional lymph node status (N stage) 3. Tumour size 4. Nuclear grade 5. Histological tumour necrosis	Score range 0–11 Low risk: score 0–2 Intermediate risk: score 3–5 High risk: score ≥6	Y

Table 1 (continued)

Risk model	Country of development	Development population	Original outcome	Risk factors/prognostic factors included	Risk groups/prognostic groups	Risk factors available
mGPS [30]	UK	Colorectal cancer		1. CRP 2. Albumin	Score Elevated CRP (>10 mg/L) and hypoalbuminaemia (<35 g/L) = 2 Elevated CRP (>10 mg/L) and albumin ≥35 g/L = 1 CRP ≤10 mg/L = 0	Potentially
PNI [41]	Japan	GI cancer	Postoperative complications	1. Serum albumin level 2. Total lymphocytes count	PNI = (10 × serum albumin level [g/100 mL]) + (0.005 × total lymphocyte count/mm^3 peripheral blood) High risk of postoperative complications if PNI ≤45 High risk of mortality if PNI <40	Potentially
Recurrence score [24]	USA	ccRCC	RFS	16 genes (11 cancer-related and 5 reference genes)	Score range 0–100 Low risk: recurrence score <32 Intermediate risk: recurrence score 32–44 High risk: recurrence score >44	N
Sao Paulo [35]	Brazil	RCC	CCS, RFS	1. Tumour size 2. Tumour grade 3. MVI	Low risk: low grade (1 or 2), diameter ≤7 cm, MVI absent Intermediate risk: 1 or 2 high risk variables High risk: high grade (3 or 4), diameter >7 cm, MVI present	Y
Sorbellini et al. [25]	USA	ccRCC	RFS	1. 2002 TNM stage 2. Tumour size (cm) 3. Fuhrman grade 4. Necrosis 5. Vascular invasion 6. Clinical presentation (a) Incidental asymptomatic, (b) Locally symptomatic (c) Systemically symptomatic	Nomogram giving continuous quantification of risk	Y
SSIGN [28]	USA	ccRCC	CSS	1. T stage 2. N stage 3. M stage 4. Tumour size 5. Nuclear grade 6. Histological tumour necrosis	Score range 0–15 Increasing score associated with decreasing CSS	Y
S-TRAC trial [37]	99 centres in 21 countries	ccRCC	RFS	1. Pathological tumour stage 2. Local nodal involvement 3. Fuhrman grade 4. ECOG-PS score	Higher risk: those with a stage 3 tumour, no or undetermined nodal involvement, no metastasis, Fuhrman grade 2 or higher, and an ECOG score of 1 or higher or a stage 4 tumour, local nodal involvement, or both	Y
TNM	UICC/AJCC	RCC	Extent of cancer spread	1. Pathological tumour stage (size of primary tumour) 2. Pathological lymph node involvement 3. Presence of metastasis	Tumours are given an overall stage based on these three risk factors which summarises the size and spread of the tumour, and thus can be used to inform management. 2002/2010/2016	Y

Table 1 (continued)

Risk model	Country of development	Development population	Original outcome	Risk factors/prognostic factors included	Risk groups/prognostic groups	Risk factors available
UISS [32]	USA	RCC	OS	1. 1997 TNM stage 2. Fuhrman grade 3. ECOG PS	Five survival stratification groups (higher group number associated with worse survival) Group I: TNM stage 1, Fuhrman Grade 1–2, PS 0 Group II: Any other TNM stage 1; TNM stage 2; TNM stage 3, any Fuhrman Grade, PS 0; TNM stage 3, Fuhrman Grade 1, PS ≥1 Group III: TNM stage 3, Fuhrman Grade 2–4, PS ≥1; TNM stage 4, Fuhrman Grade 1–2, PS 0 Group IV: TNM stage 4, Fuhrman Grade 3–4, PS 0; TNM stage 4, Fuhrman Grade 1–3, PS ≥1 Group V: TNM stage 4, FG 4, PS ≥1	Y
Wei et al. [26]	China	ccRCC	RFS	1. TNM stage 2. Fuhrman grade 3. Tumour necrosis 4. Six-SNP-based classifier	Nomogram giving continuous quantification of risk	N
Yaycioglu et al. [40]	USA	RCC	RFS	1. Preoperative clinical tumour size 2. Presentation (a) Symptomatic (b) Asymptomatic	Recurrence risk (R_{rec}) = 1.55 × presentation (0–1) + 0.19 × clinical size (in cm). Low risk: R_{rec} score ≤3.0 High risk: R_{rec} score >3.0	Y
Zisman et al. [29]	USA	RCC	CCS	1. 1997 T classification 2. Fuhrman grade 3. ECOG PS	Low risk: pT1N0M0, Fuhrman Grade 1–2, PS 0; Intermediate risk: Any other N0M0 High risk: T3N0M0, Fuhrman Grade >1, PS ≥1; Any pT4N0M0	Y

AJCC, American Joint Committee on Cancer; ccRCC, clear-cell renal cell cancer; CRP, C-reactive protein; CSS, cancer-specific survival; ECOG PS, Eastern Cooperative Oncology Group performance status; GI, gastrointestinal; GRANT, Grade, Age, Nodes and Tumour; MVI, microvascular invasion; OS, overall survival; RFS, recurrence-free survival; RRF, recurrence risk formula; SSIGN, Stage, Size Grade and Necrosis; SNP, single-nucleotide polymorphism; UICC, International Union Against Cancer; UISS, UCLA Integrated Staging System; VEGF-D, vascular endothelial growth factor-D; VEGFR, vascular endothelial growth factor receptor.

most (n = 41/59) having a median follow-up of between 60 and 90 months. The discriminative performance within all the external validations for the 19 models is shown in Fig. 1. The most frequently assessed models were the Leibovich model (n = 16), the UISS model (n = 9), the Kattan model (n = 7) and the SSIGN model (n = 7). There was substantial variation in discriminative performance both between different models and between different studies assessing the same model (Fig. 1). Meta-regression with the three risk models with eight or more external validations (Leibovich, UISS and SSIGN) showed no evidence that baseline year of recruitment, duration of prediction, study event rate or proportion of ccRCC were able to explain that heterogeneity (Table S5). The high heterogeneity also persisted in subgroup analysis based on the country of the study (Europe/US or Asian).

Figure 1 does, however, show that eight models (Jeong, Karakiewicz, Kattan, Klatte, Leibovich, Recurrence, Sorbellini and Wei) have higher discrimination than others (C-statistic ≥0.75 in at least half of the external validations and none or few C-statistics <0.7). This was confirmed in multivariate meta-analysis, where direct comparisons between the models within studies is incorporated (Fig. 2A). Those eight models all had a SUCRA of ≥0.6 (Table 2 and Fig. S2). With the exception of the Karakiewicz model that was developed for CSS, all eight had been developed for RFS in RCC. Four (Sorbellini, Karakiewicz, Leibovich and Kattan) included pathological or symptom prognostic markers that are likely to be routinely available and have been validated in at least two external populations. Jeong et al. [22] was the only model to also include age. The other three included either genetic markers (the Recurrence score), single nucleotide polymorphisms [26] or molecular markers [23] not currently available in clinical practice. These three, as well as the model by Jeong et al. [22], had only been externally validated in one population.

Conversely, the two clinical models (Yaycioglu and Cindolo), the two models based on CRP and albumin (GPS and mGPS) and the TNM criteria all had comparatively poor discrimination (SUCRA 0.1 and 0.3 and reported C-statistics of 0.63–0.70 and 0.63–0.75, respectively). Additionally, despite including the same variables as the Leibovich model, the SSIGN model, which was developed for CSS, was one of the poorest performing models, with a SUCRA of 0.4 and C-statistics below 0.7 in three of the seven external validations (range 0.63–0.78).

The multivariate meta-analysis for the Europe/US and Asian subgroups are presented in Tables S8 and S9, respectively. Except for the UISS score that performed better in European/US populations, the results were similar to those for the combined population.

In addition to the discriminative performance for RFS from the date of surgery, one study [42] included assessment of the UISS model for predicting late recurrence in patients free of disease 5 years after surgery. There was no significant difference in the probability of recurrence among those patients classified as low, intermediate and high risk based on the UISS model.

Calibration

Six studies reported calibration [20,43–47]. In a Singaporean population [47], all four models assessed (Karakiewicz, Leibovich, Kattan and Sorbellini) had reasonable calibration graphically at 5 years, with maximum departure of predicted from observed outcomes of 4%, 17%, 11% and 15%, respectively. Beisland et al. [44] found no overall evidence of miscalibration for the Leibovich model over a 10-year period in patients recruited from Norway (calibration slope 0.958). The Kattan model overestimated RFS at 5 years in two Japanese populations [43] but underestimated RFS at 5 years in a French population [46]. In a US population, the Sorbellini model [45] also underestimated the actual 5-year RFS probability in patients who had a predicted 5-year RFS probability <0.8. In a contemporary UK cohort recruited between 2011 and 2014, Vasudev et al. [20] similarly found a degree of miscalibration for 5-year RFS estimated using the Leibovich model, with the Leibovich model underestimating RFS, particularly in those at higher risk of recurrence.

Estimates of Survival for Risk Groups

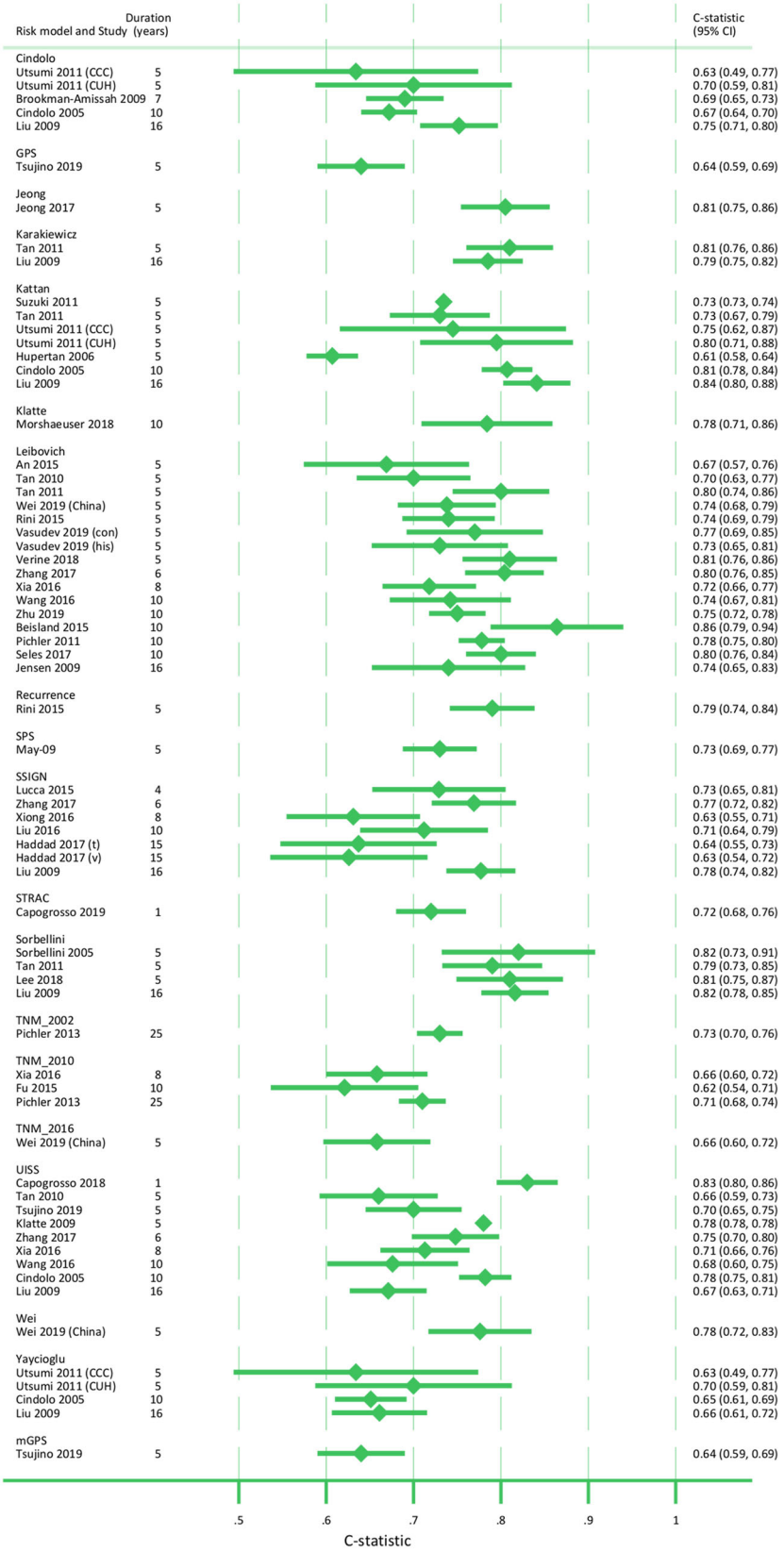
Eleven studies [20,22,55,44,48–54] reported the probability of RFS 2–10 years after surgery for risk groups determined by models (Table 3). It was not possible to pool the probabilities across studies. In all cases, the observed probability of survival decreased from the low-risk to high-risk groups.

Cancer-Specific Survival

Discrimination

Fifteen studies (Table S6) reported the discrimination of 38 external validations of 12 models for CSS from surgery. The median duration of follow-up was 33–128 months, with over half of those reporting the duration of follow-up (n = 18/34) having a median follow-up period between 60 and 90 months. As was observed for RFS there was substantial variation in the C-statistics (Fig. 3). Seven risk models, however, appeared to perform better than others, with a C-statistic of ≥0.80 in at least half of the studies in which they had been validated (Karakiewicz, Klatte, Leibovich, SSIGN, Sorbellini, Zisman and mGPS). These same seven models all had a SUCRA ≥0.6 in multivariate meta-analysis (Table 2, Fig. S3) incorporating direct comparisons (Fig. 2B).

Fig. 1 Forest plot showing the *C*-statistics from individual studies for recurrence-free survival (RFS).



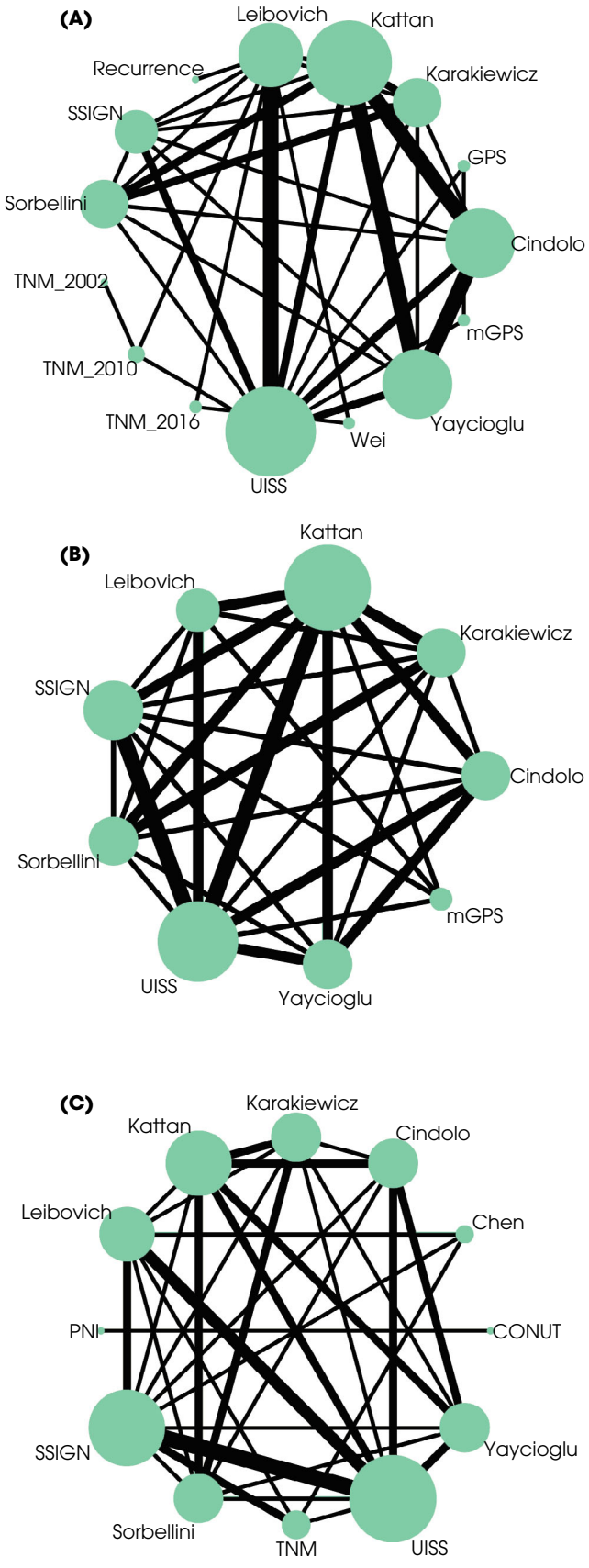


Fig. 2 Plot of direct risk model comparisons included within the multivariate meta-analysis for (**A**) Recurrence-free survival (RFS), (**B**) cancer-specific survival (CSS), and (**C**) overall survival (OS). The size of the circles and thickness of the lines are weighted according to the number of studies involved in each direct comparison. Risk models with a larger circle are therefore compared more across the studies than those with smaller circles, and risk models linked by the thickest lines are those that were most frequently compared directly against each other within the studies. CONUT, CONtrolling NUTritional status; GPS, Glasgow Prognostic Score; mGPS, modified GPS; PNI, Prognostic Nutritional Index; SSIGN, Stage, Size, Grade and Necrosis; UISS, UCLA Integrated Staging System.

Of these, the four models with the highest SUCRA (≥0.7) were the only three models developed primarily for estimating CSS (Zisman, SSIGN and Karakiewicz) and the Klatte et al. [23] model, which includes molecular markers but had only been externally validated in one cohort. The Leibovich and Sorbellini models, originally developed for RFS, were also in this group, along with the mGPS model which was originally developed for colorectal cancer prognosis and includes CRP and albumin but had also only been externally validated in one cohort.

As seen for RFS, the two models based on clinical features at presentation alone, Cindolo and Yaycioglu, had the lowest discrimination (*C*-statistics 0.65–0.71 and 0.63–0.65, respectively). Additionally, despite including the same variables as the Zisman model, the UISS model, which was developed with OS as the outcome, was one of the poorest performing models, with *C*-statistics for three of the five validations ≤0.65 and a SUCRA of 0.2. The comparative discrimination of the models was very similar when considering only European/US populations (Table S8).

In addition to the 5-year CSS from the time of surgery, Fu et al. [56] reported the performance of the SSIGN and UISS models at predicting 5-year conditional CSS, defined as the probability that a patient with RCC will survive an additional 5 years after already surviving between 1 and 5 years after surgery. The SSIGN model performed better than UISS at up to 1 year post-surgery (*C*-statistics 0.70 [0.62–0.76] and 0.65 [0.58–0.70], respectively) but there was no difference between the models from 2 to 5 years after surgery.

Calibration

Only the study by Tan et al. [47] assessed calibration. As for RFS in the same study, all four models assessed (Karakiewicz, Leibovich, Kattan and Sorbellini) had reasonable calibration graphically.

Estimates of Survival for Risk Groups

Seven studies [19,49,50,57–60] reported the probability of CSS between 1 and 10 years after surgery for risk groups determined by models (Table 3). As for RFS it was not possible to pool the probabilities across studies. In all cases,

Table 2 Multivariate meta-analysis of discrimination of risk models.

Risk model	Number of external validations	Summary risk of bias	Number of patients	Events	Borrowing of strength	Mean rank	SUCRA
Recurrence-free survival							
Jeong 2017	1	1U	93	399	0	4.5	0.8
Recurrence score	1	1U	50	1642	23.1	4.8	0.8
Sorbellini	4	3H, 1U	312	2817	22.7	4.7	0.8
Wei 2009	1	1U	98	410	23.2	5.7	0.7
Karakiewicz	2	1H, 1U	254	1043	34.7	6.1	0.7
Klatte 2009	1	1H	–	343	0	6.3	0.7
Leibovich	16	7H, 8U, 1L	1481†	7897	8.7	7.1	0.7
Kattan	7	6H, 1U	615	2851	15.7	8.2	0.6
Sao Paulo	1	1H	173	771	0	10.2	0.5
UISS	9	5H, 3U, 1L	667†	5167	17.7	10.3	0.5
S-TRAC trial	1	1H	–	730	0	10.6	0.5
TNM 2002	1	1H	443	2127	16.3	10.6	0.5
SSIGN	7	4H, 2U, 1L	542	2552	14.4	12.1	0.4
TNM 2016	1	1U	98	410	23.3	14.1	0.3
Cindolo	5	5H	532	2456	21.0	14.2	0.3
TNM 2010	3	1H, 1U, 1L	576	2580	13.1	13.9	0.3
mGPS	1	1H	–	627	22.6	15.1	0.2
GPS	1	1H	–	627	22.6	15.2	0.2
Yaycioglu	4	4H	359	1685	27.6	16.5	0.1
Cancer-specific survival							
Zisman	3	3U	1060	276	0	3.0	0.8
SSIGN	6	3H, 2U, 1L	2628	564	12.2	4.5	0.7
Karakiewicz	3	2H, 1U	1608	218	22.0	4.6	0.7
Klatte 2009	1	1H	343	–	0	4.8	0.7
Leibovich	4	3H, 1U	1524	182	17.8	5.0	0.6
mGPS	1	1H	169	35	36.6	5.3	0.6
Sorbellini	2	1H, 1U	975	174	29.2	4.9	0.6
Kattan	4	3H, 1U	3616	581	19.1	7.0	0.5
Sao Paulo	1	1H	771	122	0	8.5	0.3
UISS	6	5H, 1L	4209	659	12.6	9.9	0.2
Cindolo	2	2H	3057	483	25.8	9.7	0.2
Yaycioglu	2	2H	3057	483	24.7	10.8	0.1
Overall survival							
Chen 2017	1	1H	176	23	34.7	1.2	1
Leibovich	6	2H, 4U	1897	394	15.3	4.9	0.7
Karakiewicz	2	1H, 1U	1043	209	27.6	4.7	0.7
Sorbellini	2	1H, 1U	975	193	27.7	5.0	0.7
SSIGN	6	4H, 2U	2034	429	17.9	5.5	0.6
CONUT	1	1H	325	39	0	6.5	0.5
Kattan	3	2H, 1U	3447	750	17.9	6.5	0.5
PNI	1	1H	325	39	0	8.1	0.4
mGPS	1	1H	268	50	0	8.4	0.4
TNM (2010)	3	2H, 1U	442	118	20.8	8.5	0.4
UISS	7	3H, 4U	4622	1022	10.2	9.2	0.3
Cindolo	2	2H	3057	664	22.9	10.3	0.2
Yaycioglu	2	2H	3057	664	22.9	12.3	0.1
GRANT*	1	1H	73 217	10 059	–	–	–

*CONUT, CONtrolling NUTritional status; GPS, Glasgow Prognostic Score; GRANT, Grade, Age, Nodes and Tumour; H, high risk of bias, L, low risk of bias; mGPS, modified GPS; PNI, Prognostic Nutritional Index; SSIGN, Stage, Size Grade and Necrosis; U, unclear risk of bias; UISS, UCLA Integrated Staging System. *Excluded from multivariate analysis as only assessed in one population with no other risk models. †Events not reported for two studies.*

the observed probability of survival decreased moving from the low-risk to the high-risk groups.

Overall Survival

Discrimination

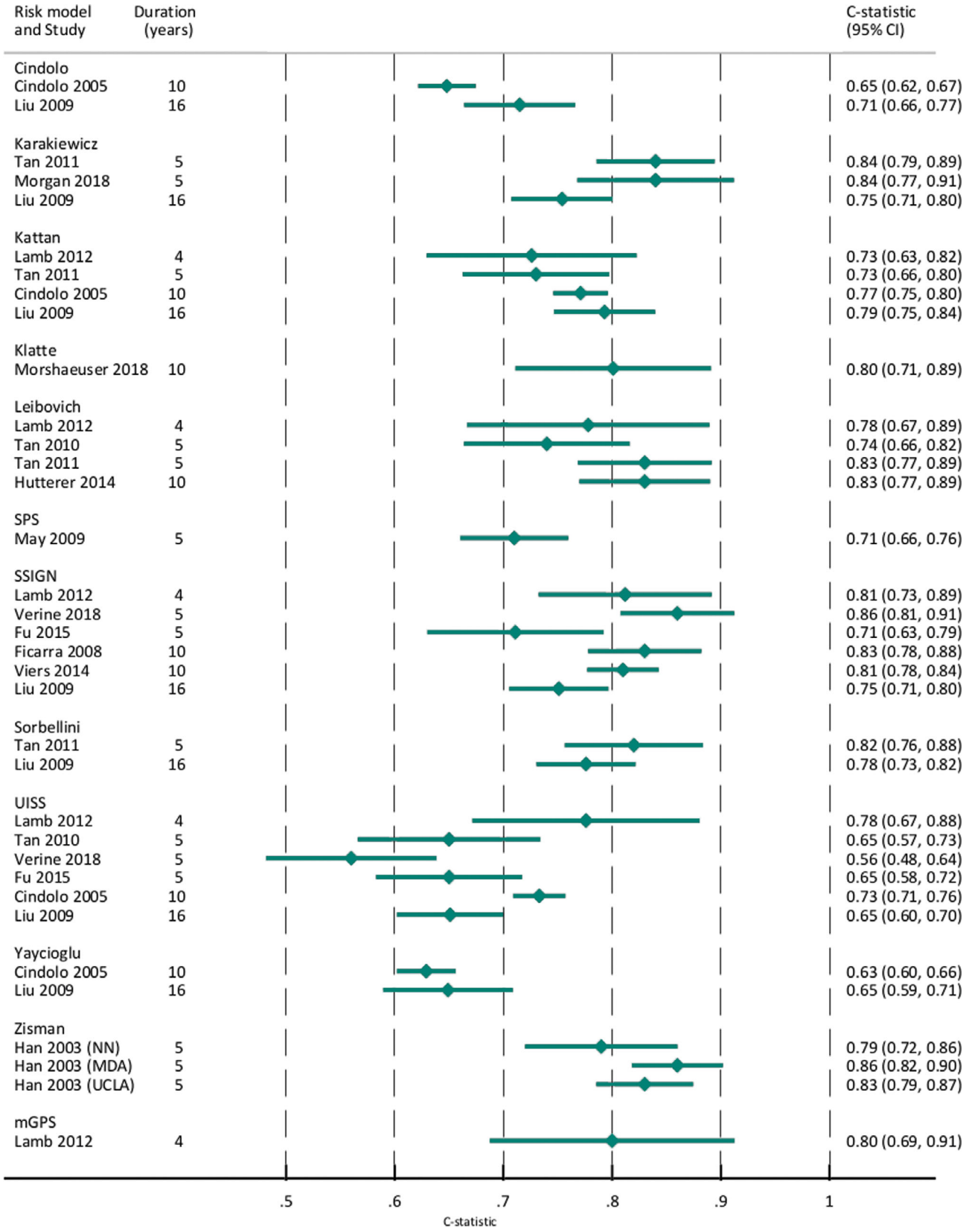
Twenty studies (Table S7) reported the discrimination of 43 external validations of models for OS. As for RFS and CSS, heterogeneity was high (I^2 up to 87.5%), therefore, estimates were not pooled. The reported *C*-statistics for different models ranged from 0.59 to 0.90 (Fig. 4). The model with the highest discrimination in any validation (*C*-statistic 0.90 [0.80–0.95]) was the model developed by Chen et al., which includes pathological T-stage along with three biochemical ratios. That model, however, had only been validated in one small population (23 cases of RCC)

Table 3 Estimates of the probability of survival for subgroups based on different prognostic models.

Risk model	Time period, years	Probability of survival			Study	Country	Recruitment period	Overall risk of bias
		Low risk/good prognosis	Intermediate risk/prognosis	High risk/poor prognosis				
Recurrence-free survival								
Cindolo	2	0.92	–	0.79	Brookman-Amissah 2009 [48]	Germany	1992–2006	High
Sao Paolo	5	0.91	0.61	0.52	May 2009 [49]	Germany	1992–2006	High
Cindolo	5	0.85	–	0.68	Brookman-Amissah 2009 [48]	Germany	1992–2006	High
CONUT	5	0.87	–	0.59	Song 2019 [50]	China	2010–2012	High
Jeong 2017	5	0.95	0.64	0.34	Jeong 2017 [22]	South Korea	2005–2011	Unclear
Leibovich	5	0.89	0.70	0.44	Xu 2015 [51]	China	2001–2004	Unclear
		0.88	0.68	0.35	Jensen 2009 [52]	Denmark	1992–2001	Unclear
		0.97	0.85	0.5	Vasudev 2019 [20]	UK	2011–2014	High
		0.93	0.76	0.37	Vasudev 2019 [20]	UK	1998–2006	High
UISS	5	0.88	0.72	0.50	Xu 2015 [51]	China	2001–2004	Unclear
		0.91	0.78	0.53	Chang 2015 [53]	China	2003–2004	Unclear
Cindolo	7	0.81	–	0.56	Brookman-Amissah 2009 [48]	Germany	1992–2006	High
Leibovich	10	0.91	0.71	0.26	Picher 2011 [54]	Austria	1984–2006	Unclear
		0.87	0.64	0.20	Beisland 2015 [44]	Norway	1997–2013	High
		0.95	0.87	0.64	Seles 2017* [55]	Austria	2005–2013	High
Cancer-specific survival								
Zisman	1	0.98	0.91	0.73	Han 2003 (NN) [57]	The Netherlands	1990–2001	Unclear
		0.98	0.97	0.80	Han 2003 (MDA) [57]	USA	1987–2000	Unclear
		1.00	0.97	0.81	Han 2003 (UCLA) [57]	USA	1989–2001	Unclear
mGPS	2	0.99	0.73	0.44	Tsuijino 2018 [58]	Japan	2005–2015	High
Zisman	3	0.94	0.77	0.44	Han 2003 (NN) [57]	The Netherlands	1990–2001	Unclear
		0.98	0.85	0.52	Han 2003 (MDA) [57]	USA	1987–2000	Unclear
		0.95	0.87	0.58	Han 2003 (UCLA) [57]	USA	1989–2001	Unclear
mGPS (0,1,2)	4	0.96	0.74	0.00	Lamb 2012 [19]	UK	1997–2007	High
CONUT	5	0.95	–	0.73	Song 2019 [50]	China	2010–2012	High
Karakiewicz	5	0.99	–	0.84	Morgan 2018 [59]	USA	2000–2009	High
Sao Paolo	5	0.94	0.80	0.59	May 2009 [49]	Germany	1992–2006	High
Zisman	5	0.94	0.65	0.40	Han 2003 (NN) [57]	The Netherlands	1990–2001	Unclear
		0.92	0.73	0.30	Han 2003 (MDA) [57]	USA	1987–2000	Unclear
		0.93	0.78	0.48	Han 2003 (UCLA) [57]	USA	1989–2001	Unclear
UISS	10	0.62	0.73	1.00	Ficarra 220 [60]	Italy	1986–2000	Unclear
Overall survival								
UISS	1	0.99	0.95	0.82	Cindolo 2008 [62]	Italy, France and Austria	1984–2002	High
mGPS	2	0.98	0.73	0.44	Tsujino 2018 [58]	Japan	2005–2015	High
UISS	2	0.97	0.89	0.74	Cindolo 2008 [62]	Italy, France and Austria	1984–2002	High
UISS	3	0.96	0.84	0.65	Cindolo 2008 [62]	Italy, France and Austria	1984–2002	High
UISS	4	0.93	0.79	0.58	Cindolo 2008 [62]	Italy, France and Austria	1984–2002	High
CONUT	5	0.94	–	0.68	Song 2019 [50]	China	2010–2012	High
GRANT	5	0.94	0.86/0.76	0.46	Buti 2019 [63]	USA	2001–2015	High
UISS	5	0.94	0.88	0.73	Chang 2015 [53]	China	2003–2004	Unclear
		0.90	0.74	0.52	Cindolo 2008 [62]	Italy, France and Austria	1984–2002	High

*CONUT, CONtrolling NUTritional status; GRANT, Grade, Age, Nodes and Tumour; mGPS, modified Glasgow Prognostic Score; UISS, UCLA Integrated Staging System. *Although median follow-up only 6.1 years.*

Fig. 3 Forest plot showing the *C*-statistics from individual studies for cancer-specific survival (CSS).



from the same hospital at which the model was developed. The model may not perform as well in other populations.

As seen for RFS and CSS, the two models based on clinical features at presentation alone, Cindolo and Yaycioglu, had the lowest discrimination (*C*-statistics 0.62–0.70 and 0.59–0.62, respectively). Despite being developed for OS, the UISS model also had comparatively low discrimination, with a *C*-statistic of <0.7 in four of the six validation studies and *C*-statistics consistently lower than those for Leibovich, SSIGN, Karakiewicz and Sorbellini in direct comparisons. This was reflected in the multivariate analysis where the UISS model had a SUCRA of 0.3 and, together with the Chen et al. model, the four highest ranking models with SUCRA values ≥6 were the Leibovich, SSIGN, Karakiewicz and Sorbellini models. There was little to distinguish among those four, with all the models also including pathological or symptomatic prognostic factors likely to be routinely available in clinical practice. The comparative discrimination of the models was very similar when considering only Asian populations (Table S9).

Calibration

Two studies reported data on calibration. As for RFS and CSS, the study by Tan et al. reported that all four models assessed (Karakiewicz, Leibovich, Kattan and Sorbellini) had reasonable calibration graphically [47]. Using the 'validation by calibration' approach [61], Cindolo et al. [62] found that the UISS model significantly (likelihood ratio test $P < 0.0001$) underestimated OS, particularly at the extremes. The difference was mainly attributable to a population-level underestimation bias, with no evidence that the relative effects of the risk factors in the model were inadequately estimated.

Estimates of Survival for Risk Groups

Five studies [50,53,58,62,63] reported the probability of OS between 1 and 5 years after surgery for risk groups determined by models (Table 3). As for RFS and CSS, it was not possible to pool the probabilities across studies and in all cases the probability of survival fell when moving from low-risk to high-risk groups.

Improvement in Performance of Previously Published Risk Models with the Addition of Additional Prognostic Markers

Forty studies externally validated pre-existing risk models and also investigated the improvement in the performance of these models when additional prognostic markers were incorporated (Table S10). Thirty-five studies evaluated additional prognostic markers for RFS, three for CSS and 15 for OS. Improvements in the *C*-statistic of up to 0.171 were observed. However, of the 40 additional prognostic markers, 28 required assessment using immunohistochemistry, *in situ* hybridization, or quantitative RT-PCR not currently routinely available in clinical practice.

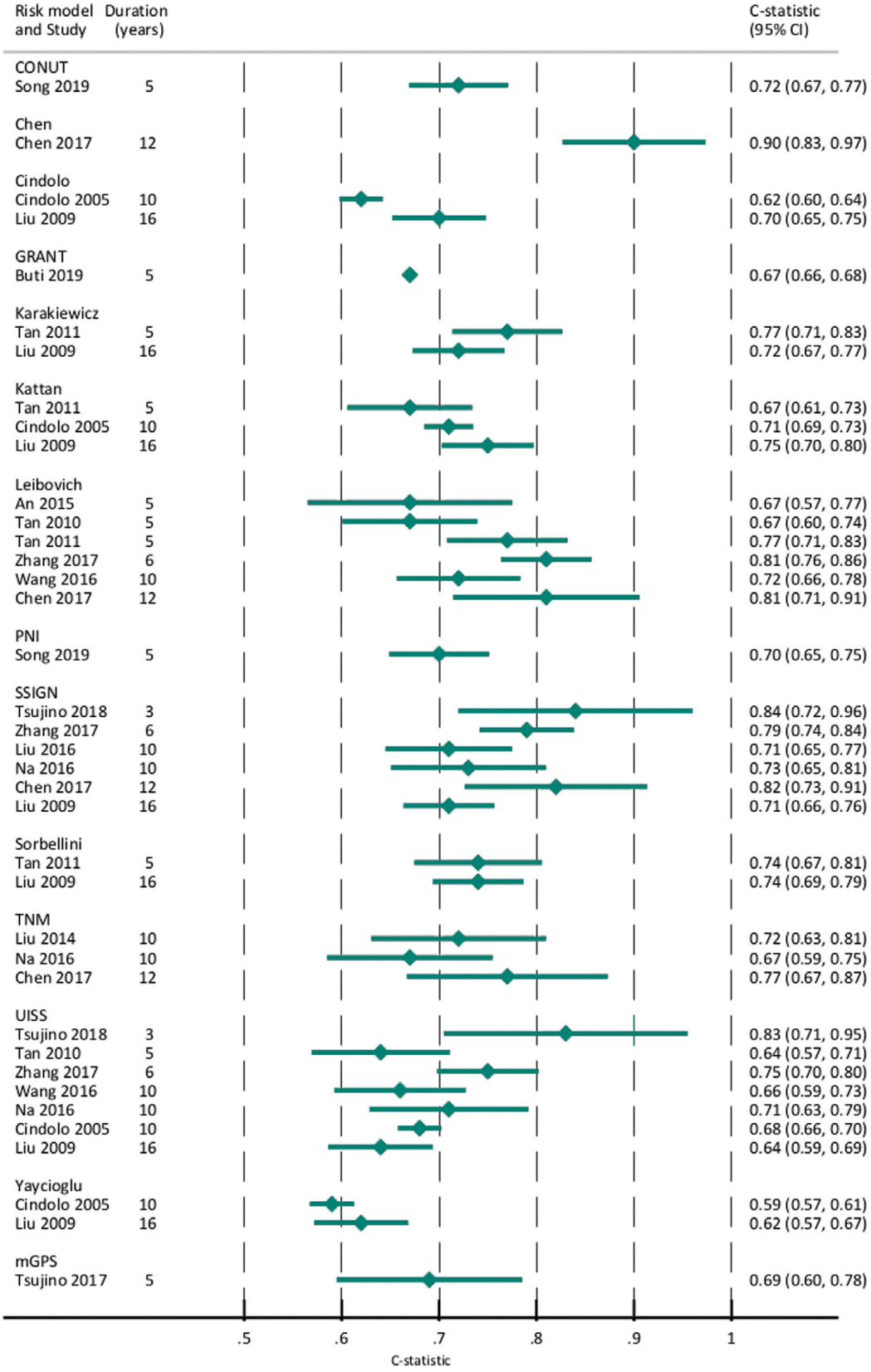
Discussion

This review shows that there is no clear single 'best' model for any of the three outcomes considered (RFS, CSS and OS). Instead, there are several risk models that have all been assessed in at least two external populations and have similarly high discriminative performance. For RFS, these are the Sorbellini, Karakiewicz, Leibovich and Kattan models, with UISS also having comparable performance in European/US populations. For CSS, they are the Zisman, SSIGN, Karakiewicz, Leibovich and Sorbellini models, and for OS they are the Leibovich, Karakiewicz, Sorbellini and SSIGN models. All performed better than TNM alone. Ideally the choice between these models for a given setting would be based on validation studies in the relevant population of interest [9]. This review provides the most comprehensive summary to date of the performance of the models in different populations. Where data are not currently available for a specific population or several models remain similar, the choice should depend on the availability and accuracy of data on the prognostic factors included in each risk model. For example, from the six better-performing models across the three outcomes, the Leibovich and SSIGN models require only routinely reported tumour pathology data, while the Karakiewicz, Sorbellini and Kattan models include symptoms at presentation and the Zisman model includes Eastern Cooperative Oncology Group performance status. Three models (Sorbellini, Karakiewicz and Leibovich) also ranked highly for all three survival outcomes so, if a prognostic model is to be used to predict all three, one of those models would be most appropriate.

In addition to these six models, there were also several models that had similar performance but had only been assessed in one external population so further validation studies are required. These include models which use genetic risk markers (Recurrence score) [26], molecular markers (Klatte), biochemical markers [21] (mGPS) and age [22]. While these models have limited current clinical utility within routine practice, they may be of utility in the future or within clinical trials.

This review additionally shows that there are some models that are unlikely to be the most appropriate choice in any setting. Of particular note, the SSIGN model cited in the ESMO guidelines performed comparatively poorly for RFS, and the UISS model, highlighted in both the ESMO and EAU guidelines, is unlikely to be best choice for either CSS or OS.

Fig. 4 Forest plot showing the *C*-statistics from individual studies for overall survival (OS).



While estimates of model calibration were only infrequently included, most models that were assessed underestimated survival, particularly in more recent populations. As discussed elsewhere [20] this may be due to improvements in imaging and surgical techniques. If the models are to be used to provide individualized estimates to patients or to compare RCC outcomes with competing health risks, all would need recalibrating to the specific setting.

A key strength of this review was our systematic search of multiple databases, enabling us to identify more models and more external validations than previous reviews [5,6]. Although the heterogeneity of the included studies limited the pooling of data, our use of multivariate meta-analysis techniques enabled us to rank the relative discrimination of the models. This approach incorporates both direct and indirect comparisons and so takes into account the relative performance of risk models within individual studies and limits the effects of heterogeneity among the studies. It does, however, assume that the relative performance of risk models in one study is transferable to other studies and that missing comparisons are missing at random. These assumptions are unlikely to be true in all cases owing to selective outcome reporting [64] or to selective choice of analyses [65]. Most of the included studies were also at moderate or high risk of bias and the small number of studies at low risk of bias meant it was not possible to perform a subgroup analysis including only those studies. All but two of the included studies also evaluated the performance of models in retrospective cohorts. These studies are at risk of both collection and ascertainment bias through a lack of standardization over data collection, potential differences in reporting and collection methods both between centres and over time, and a lack of centralized pathological review. The recruitment periods of many of the studies also began more than 20 years ago and so the outcomes may not reflect current practice. The biggest change in clinical care over that time, the shift from routine open partial nephrectomy to robot-assisted partial nephrectomy, however, is unlikely to have significantly impacted on survival estimates as current data suggest that there are no differences in oncological outcomes after open partial nephrectomy, laparoscopy partial nephrectomy or robot-assisted partial nephrectomy [66–68]. However, further validation in contemporary cohorts, ideally from large prospective studies, are needed.

Reflecting their intended use in clinical practice, most models were also assessed as scores rather than using the original model coefficients. By including only those models that had been externally validated, we have also not included more recent models that are yet to be assessed externally, for example, the D-SSIGN adaptation of the SSIGN model developed for dynamic risk prediction [69], the RCC histology-specific Leibovich models [70] and a new model developed in the ASSURE trial population for patients with high-risk localized and locally advanced RCC [71]. Although our decision to only include external validation studies in unselected cohorts of patients presenting with RCC or ccRCC means that our findings reflect the performance of the risk models in routine clinical practice, we note that the performance metrics may differ within select groups, such as those considered at high risk and recruited to adjuvant clinical trials. As seen in a recent validation [71], the discrimination is likely to be poorer in such populations where the case mix is narrower due to the prior exclusion of those at low risk [72].

In summary, this review shows that there are at least six prognostic models that include data available within routine clinical practice and that have better discriminative ability than TNM staging alone for RFS, CSS and OS in patients treated with surgery for localized ccRCC. This supports current EAU and ESMO guideline recommendations to use prognostic models to inform surveillance, while also confirming that there is currently no single 'best' model. The findings on the comparative performance and the prognostic factors included in the models in this review should support clinicians and guideline developers to make an informed choice of which model to use for current surveillance. Additionally, in light of recent promising data from adjuvant trials [7], the findings are likely to be of increasing importance. As highlighted recently [73], all of the 11 largest RCC adjuvant trials that have completed or are currently recruiting rely on one or more prognostic models to determine eligibility. Selection of the most appropriate prognostic model is therefore important not only for the design and recruitment of future clinical trials but also for decisions on who may or may not be offered adjuvant treatment. Given the significant potential harms associated with adjuvant treatment, prognostic models will be a key resource for supporting informed decision making with patients. All would need recalibration if individualized risk estimates of outcomes are used.

Acknowledgements

We thank Isla Kuhn for help developing the search strategy and our two patient and public representatives Tim Cribb and Dave Ellwood for their input and advice during this study and comments on this manuscript. We also thank Sarah Norman for administrative support throughout the project.

Juliet A. Usher-Smith was funded by a Cancer Research UK Prevention Fellowship (C55650/A21464). The University of Cambridge has received salary support in respect of Simon J. Griffin from the NHS in the East of England through the Clinical Academic Reserve. Sabrina H. Rossi is funded by a Cancer Research UK Clinical PhD Fellowship. Grant D. Stewart is supported by the Renal Cancer Research Fund, the

Mark Foundation for Cancer Research, the Cancer Research UK Cambridge Centre [C9685/A25177] and the National Institute of Health Research (NIHR) Cambridge Biomedical Research Centre (BRC-1215-20014). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. Hannah Harrison was supported by an NIHR Methods Fellowship (RM-SR-2017-09-009) and is now supported by a NIHR Development and Skills Enhancement Award (NIHR301182). Stephen J. Sharp is funded by the Medical Research Council (MC_UU_00006/6). The funders had no role in the design and conduct of the study, collection, management, analysis and interpretation of the data, or preparation, review or approval of the manuscript. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Conflict of Interest

Grant D. Stewart has received educational grants from Pfizer, AstraZeneca and Intuitive Surgical, consultancy fees from Pfizer, Merck, EUSA Pharma and CMR Surgical, Travel expenses from Pfizer and Speaker fees from Pfizer. All other authors have no financial disclosures.

Data Availability Statement

All data used in this study are publicly available in the primary articles. Juliet A. Usher-Smith had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

References

1 **Donat SM, Diaz M, Bishoff JT et al**. American Urological Association (AUA) guideline: follow-up for clinically localized renal neoplasms. *AUA Clin Guidel* 2013; 190: 407–16

2 **National Comprehensive Cancer Network**. *NCCN Clinical Practice Guidelines in Oncology: Kidney Cancer*. National Comprehensive Cancer Network, 2021. Available at: https://www.nccn.org/professionals/physician_gls/pdf/kidney.pdf. Accessed January 4, 2022.

3 **Escudier B, Porta C, Schmidinger M et al**. Renal cell carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2019; 30: 706–20

4 **Ljungberg B, Albiges L, Bedke J et al**. EAU guidelines on renal cell carcinoma [Internet], 2021. Available at: https://uroweb.org/guideline/renal-cell-carcinoma/. Accessed August 2021

5 **Sun M, Shariat SF, Cheng C et al**. Prognostic factors and predictive models in renal cell carcinoma: a contemporary review. *Eur Urol* 2011; 60: 644–61

6 **Klatte T, Rossi SH, Stewart GD**. Prognostic factors and prognostic models for renal cell carcinoma: a literature review. *World J Urol* 2018; 36: 1943–52

7 **Choueiri TK, Tomczak P, Park SH et al**. Pembrolizumab versus placebo as post-nephrectomy adjuvant therapy for patients with renal cell carcinoma: randomized, double-blind, phase III KEYNOTE-564 study. *Am Soc Clin Oncol Annu Meet* 2021; 39(18_suppl): LBA5

8 **Debray TPA, Damen JAAG, Snell KIE et al**. A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017; 356: i6460.

9 **Collins GS, Reitsma JB, Altman DG, Moons KGM**. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015; 162: 55–63

10 **Kramar A, Negrier S, Sylvester R et al**. Guidelines for the definition of time-to-event end points in renal cell cancer clinical trials: results of the DATECAN project. *Ann Oncol* 2015; 26: 2392–8

11 **Wolff RF, Moons KGM, Riley RD et al**. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2018; 170: 51

12 **Higgins JP, Thompson SG, Deeks JJ, Altman DG**. Measuring inconsistency in meta-analyses. *BMJ* 2003; 327: 557–60

13 **Snell KIE, Ensor J, Debray TPA, Moons KGM, Riley RD**. Meta-analysis of prediction model performance across multiple studies: which scale helps ensure between-study normality for the C-statistic and calibration measures? *Stat Methods Med Res* 2018; 27: 3505–22

14 **White IR**. Multivariate random-effects meta-regression: updates to mvmeta. *Stata J* 2011; 11: 255–70

15 **White IR**. Multivariate random-effects meta-analysis. *Stata J* 2009; 9: 40–56

16 **Riley RD, Thompson JR, Abrams KR**. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics* 2008; 9: 172–86

17 **Jackson D, White IR, Price M, Copas J, Riley RD**. Borrowing of strength and study weights in multivariate and network meta-analysis. *Stat Methods Med Res* 2017; 26: 2853–68

18 **Salanti G, Ades AE, Ioannidis JPA**. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol* 2011; 64: 163–71

19 **Lamb GWA, Aitchison M, Ramsey S, Housley SL, McMillan DC**. Clinical utility of the Glasgow Prognostic Score in patients undergoing curative nephrectomy for renal clear cell cancer: Basis of new prognostic scoring systems. *Br J Cancer* 2012; 106: 279–83

20 **Vasudev NS, Hutchinson M, Trainor S et al**. UK multicenter prospective evaluation of the Leibovich score in localized renal cell carcinoma: performance has altered over time. *Urology* 2020; 136: 162–8

21 **Chen Z, Shao Y, Yao H et al**. Preoperative albumin to globulin ratio predicts survival in clear cell renal cell carcinoma patients. *Oncotarget* 2017; 8: 48291–302

22 **Jeong SU, Park J-M, Shin S-J et al**. Prognostic significance of macroscopic appearance in clear cell renal cell carcinoma and its metastasis-predicting model. *Pathol Int* 2017; 67: 610–9

23 **Klatte T, Seligson DB, LaRochelle J et al**. Molecular signatures of localized clear cell renal cell carcinoma to predict disease-free survival after nephrectomy. *Cancer Epidemiol Biomarkers Prev* 2009; 18: 894–900

24 **Rini B, Goddard A, Knezevic D et al**. A 16-gene assay to predict recurrence after surgery in localised renal cell carcinoma: development and validation studies. *Lancet Oncol* 2015; 16: 676–85

25 **Sorbellini M, Kattan MW, Snyder ME et al**. A postoperative prognostic nomogram predicting recurrence for patients with conventional clear cell renal cell carcinoma. *J Urol* 2005; 173: 48–51

26 **Wei JH, Feng ZH, Cao Y et al**. Predictive value of single-nucleotide polymorphism signature for recurrence in localised renal cell carcinoma: a retrospective analysis and multicentre validation study. *Lancet Oncol* 2019; 20: 591–600

27 **Kattan M, Reuter V, Motzer R, Katz J, Russo P**. A postoperative prognostic nomogram for renal cell carcinoma. *J Urol* 2001; 166: 63–7

28 **Frank I, Blute M, Cheville J, Lohse C, Weaver A, Zincke H**. An outcome prediction model for patients with clear cell renal cell carcinoma treated with radical nephrectomy based on tumor stage, size, grade and necrosis: the SSIGN score. *J Urol J Urol* 2002; 168: 2395–400

29 **Zisman A**, **Pantuck AJ**, **Wieder J et al**. Risk group assessment and clinical outcome algorithm to predict the natural history of patients with surgically resected renal cell carcinoma. *J Clin Oncol* 2002; 20: 4559–66

30 **McMillan DC**, **Crozier JEM**, **Canna K**, **Angerson WJ**, **McArdle CS**. Evaluation of an inflammation-based prognostic score (GPS) in patients undergoing resection for colon and rectal cancer. *Int J Colorectal Dis* 2007; 22: 881–6

31 **Cindolo L**, **de la Taille A**, **Messina G et al**. A preoperative clinical prognostic model for non-metastatic renal cell carcinoma. *BJU Int* 2003; 92: 901–5

32 **Zisman A**, **Pantuck AJ**, **Dorey F et al**. Improved prognostication of renal cell carcinoma using an integrated staging system. *J Clin Oncol* 2001; 19: 1649–57

33 **Buti S**, **Puligandla M**, **Bersanelli M et al**. Validation of a new prognostic model to easily predict outcome in renal cell carcinoma: the GRANT score applied to the ASSURE trial population. *Ann Oncol* 2017; 28: 2747–53

34 **Karakiewicz PI**, **Briganti A**, **Chun F-H et al**. Multi-institutional validation of a new renal cancer-specific survival nomogram. *J Clin Oncol* 2007; 25: 1316–22

35 **Dall'Oglio MF**, **Ribeiro-Filho LA**, **Antunes AA et al**. Microvascular tumor invasion, tumor size and Fuhrman grade: a pathological triad for prognostic evaluation of renal cell carcinoma. *J Urol* 2007; 178: 425–8

36 **Ignacio de Ulibarri J**, **Gonzalez-Madrono A**, **de Villar N et al**. CONUT: a tool for controlling nutritional status. First validation in a hospital population. *Nutr Hosp* 2005; 20: 38–45

37 **Ravaud A**, **Motzer RJ**, **Pandha HS et al**. Adjuvant sunitinib in high-risk renal-cell carcinoma after nephrectomy. *N Engl J Med* 2016; 375: 2246–54

38 **Leibovich BC**, **Blute ML**, **Cheville JC et al**. Prediction of progression after radical nephrectomy for patients with clear cell renal cell carcinoma: a stratification tool for prospective clinical trials. *Cancer* 2003; 97: 1663–71

39 **Forrest LM**, **McMillan DC**, **McArdle CS**, **Angerson WJ**, **Dunlop DJ**. Evaluation of cumulative prognostic scores based on the systemic inflammatory response in patients with inoperable non-small-cell lung cancer. *Br J Cancer* 2003; 89: 1028–30

40 **Yaycioglu O**, **Roberts WW**, **Chan T**, **Epstein JI**, **Marshall FF**, **Kavoussi LR**. Prognostic assessment of nonmetastatic renal cell carcinoma: a clinically based model. *Urology* 2001; 58: 141–5

41 **Onodera T**, **Goseki N**, **Kosaki G**. Prognostic nutritional index in gastrointestinal surgery of malnourished cancer patients. *Nihon Geka Gakkai Zasshi* 1984; 85: 1001–5

42 **Capogrosso P**, **Larcher A**, **Sjoberg DD et al**. Risk-based surveillance after surgical treatment of renal cell carcinoma. *J Urol* 2018; 200: 61–7

43 **Utsumi T**, **Ueda T**, **Fukasawa S et al**. Prognostic models for renal cell carcinoma recurrence: external validation in a Japanese population. *Int J Urol* 2011; 18: 667–71

44 **Beisland C**, **Gudbrandsdottir G**, **Reisæter LAR**, **Bostad L**, **Wentzel-Larsen T**, **Hjelle KM**. Contemporary external validation of the Leibovich model for prediction of progression after radical surgery for clear cell renal cell carcinoma. *Scand J Urol* 2015; 49: 205–10

45 **Lee BH**, **Feifer A**, **Feuerstein MA et al**. Validation of a postoperative nomogram predicting recurrence in patients with conventional clear cell renal cell carcinoma. *Eur Urol Focus* 2018; 4: 100–5

46 **Hupertan V**, **Roupret M**, **Poisson J-F et al**. Low predictive accuracy of the Kattan postoperative nomogram for renal cell carcinoma recurrence in a population of French patients. *Cancer* 2006; 107: 2604–8

47 **Tan M-H**, **Li H**, **Choong CV et al**. The Karakiewicz nomogram is the most useful clinical predictor for survival outcomes in patients with localized renal cell carcinoma. *Cancer* 2011; 117: 5314–24

48 **Brookman-Amissah S**, **Kendel F**, **Spivak I et al**. Impact of clinical variables on predicting disease-free survival of patients with surgically resected renal cell carcinoma. *BJU Int* 2009; 103: 1375–80

49 **May M**, **Brookman-Amissah S**, **Kendel F et al**. Validation of a postoperative prognostic model consisting of tumor microvascular invasion, size, and grade to predict disease-free and cancer-specific survival of patients with surgically resected renal cell carcinoma. Original article: Clinical investiga. *Int J Urol* 2009; 16: 616–21

50 **Song H**, **Xu B**, **Luo C et al**. The prognostic value of preoperative controlling nutritional status score in non-metastatic renal cell carcinoma treated with surgery: a retrospective single-institution study. *Cancer Manag Res* 2019; 11: 7567–75

51 **Xu L**, **Zhu Y**, **An H et al**. Clinical significance of tumor-derived IL-1β and IL-18 in localized renal cell carcinoma: associations with recurrence and survival. *Urol Oncol* 2015; 33: 68.e9–16

52 **Jensen HK**, **Donskov F**, **Marcussen N**, **Nordsmark M**, **Lundbeck F**, **Von Der Maase H**. Presence of intratumoral neutrophils is an independent prognostic factor in localized renal cell carcinoma. *J Clin Oncol* 2009; 27: 4709–17

53 **Chang Y**, **Xu L**, **An H et al**. Expression of IL-4 and IL-13 predicts recurrence and survival in localized clear-cell renal cell carcinoma. *Int J Clin Exp Pathol* 2015; 8: 1594–603

54 **Pichler M**, **Hutterer GC**, **Chromecki TF et al**. Prognostic value of the leibovich prognosis score supplemented by vascular invasion for clear cell renal cell carcinoma. *J Urol* 2012; 187: 834–9

55 **Seles M**, **Posch F**, **Pichler GP et al**. Blood platelet volume represents a novel prognostic factor in patients with nonmetastatic renal cell carcinoma and improves the predictive ability of established prognostic scores. *J Urol* 2017; 198: 1247–52

56 **Fu Q**, **Chang Y**, **An H et al**. Prognostic value of interleukin-6 and interleukin-6 receptor in organ-confined clear-cell renal cell carcinoma: a 5-year conditional cancer-specific survival analysis. *Br J Cancer* 2015; 113: 1581–9

57 **Han K-R**, **Bleumer I**, **Pantuck AJ et al**. Validation of an integrated staging system toward improved prognostication of patients with localized renal cell carcinoma in an international population. *J Urol* 2003; 170: 2221–4

58 **Tsujino T**, **Komura K**, **Matsunaga T et al**. Preoperative measurement of the modified Glasgow prognostic score predicts patient survival in non-metastatic renal cell carcinoma prior to nephrectomy. *Ann Surg Oncol* 2017; 24: 2787–93

59 **Morgan TM**, **Mehra R**, **Tiemeny P et al**. A multigene signature based on cell cycle proliferation improves prediction of mortality within 5 yr of radical nephrectomy for renal cell carcinoma. *Eur Urol* 2018; 73: 763–9

60 **Ficarra V**, **Novara G**, **Galfano A et al**. The "Stage, Size, Grade and Necrosis" score is more accurate than the University of California Los Angeles Integrated Staging System for predicting cancer-specific survival in patients with clear cell renal cell carcinoma. *BJU Int* 2009; 103: 165–70

61 **Van Houwelingen HC**. Validation, calibration, revision and combination of prognostic survival models. *Stat Med* 2000; 19: 3401–15

62 **Cindolo L**, **Chiodini P**, **Gallo C et al**. Validation by calibration of the UCLA integrated staging system prognostic model for nonmetastatic renal cell carcinoma after nephrectomy. *Cancer* 2008; 113: 65–71

63 **Buti S**, **Karakiewicz PI**, **Bersanelli M et al**. Validation of the GRade, Age, Nodes and Tumor (GRANT) score within the Surveillance Epidemiology and End Results (SEER) database: a new tool to predict survival in surgically treated renal cell carcinoma patients. *Sci Rep* 2019; 9: 1–7

64 **Kirkham JJ**, **Dwan KM**, **Altman DG et al**. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ* 2010; 340: 637–40

65 **Dwan K**, **Altman DG**, **Clarke M et al**. Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials. Fugh-Berman AJ, editor. *PLoS Med* 2014; 11: e1001666

66 **Chang KD**, **Abdel Raheem A**, **Kim KH et al**. Functional and oncological outcomes of open, laparoscopic and robot-assisted partial nephrectomy: a

multicentre comparative matched-pair analyses with a median of 5 years' follow-up. *BJU Int* 2018; 122: 618–26

67 **Peyronnet B**, **Seisen T**, **Oger E et al**. Comparison of 1800 robotic and open partial nephrectomies for renal tumors. *Ann Surg Oncol* 2016; 23: 4277–83

68 **Choi JE**, **You JH**, **Kim DK**, **Rha KH**, **Lee SH**. Comparison of perioperative outcomes between robotic and laparoscopic partial nephrectomy: a systematic review and meta-analysis. *Eur Urol* 2015; 67: 891–901

69 **Thompson RH**, **Leibovich BC**, **Lohse CM et al**. Dynamic outcome prediction in patients with clear cell renal cell carcinoma treated with radical nephrectomy: the D-SSIGN score. *J Urol* 2007; 177: 477–80

70 **Leibovich BC**, **Lohse CM**, **Cheville JC et al**. Predicting oncologic outcomes in renal cell carcinoma after surgery. *Eur Urol* 2018; 73: 772–80

71 **Correa AF**, **Jegede OA**, **Haas NB et al**. Predicting disease recurrence, early progression, and overall survival following surgical resection for high-risk localized and locally advanced renal cell carcinoma. *Eur Urol* 2021; 80: 20–31

72 **Riley RD**, **Ensor J**, **Snell KIE et al**. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016; 353: 27–30

73 **Correa AF**, **Jegede O**, **Haas NB et al**. Predicting renal cancer recurrence: defining limitations of existing prognostic models with prospective trial-based validation. *J Clin Oncol* 2019; 37: 2062–71

Correspondence: Juliet A. Usher-Smith, The Primary Care Unit, Department of Public Health and Primary Care, University of Cambridge School of Clinical Medicine, Box 113 Cambridge Biomedical Campus, Cambridge CB2 0SR, UK.

e-mail: jau20@medschl.cam.ac.uk

Abbreviations: ccRCC, clear-cell RCC; CRP, C-reactive protein; CSS, cancer-specific survival; EAU, European Association of Urology; ESMO, European Society for Medical Oncology; GRANT, Grade, Age, Nodes and Tumour; NCCN, National Comprehensive Cancer Network; NIHR, National Institute for Health Research; OS, overall survival; RFS, recurrence-free survival; RoB, risk of bias; SSIGN, Stage, Size, Grade and Necrosis; SUCRA, surface under the cumulative ranking curve; UISS, UCLA Integrated Staging System.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Method S1. Details of risk of bias assessment.
Table S1. Medline search strategy.
Table S2. Embase search strategy.
Table S3. Characteristics of included studies.
Table S4. Key study characteristics and risk of bias assessment for external validations of models predicting recurrence-free survival.
Table S5. Results of meta-regression for recurrence-free survival.
Table S6. Key study characteristics and risk of bias assessment for external validations of models predicting cancer-specific survival.
Table S7. Key study characteristics and risk of bias assessment for external validations of models predicting overall survival.
Table S8. Multivariate meta-analysis of discrimination of risk models in Europe/US populations.
Table S9. Multivariate meta-analysis of discrimination of risk models in Asian populations.
Table S10. Discrimination of externally validated risk models without and with the addition of one or more additional prognostic markers.
Fig. S1. PRISMA flow diagram.
Fig. S2. Plots of the ranking for each risk score considered in the multivariate meta-analysis for recurrence-free survival.
Fig. S3. Plots of the ranking for each risk score considered in the multivariate meta-analysis for cancer-specific survival.
Fig. S4. Plots of the ranking for each risk score considered in the multivariate meta-analysis for overall survival.