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

Juliet A. Usher-Smith[a]*, Lanxin Li[b]*, Lydia Roberts[c]*, Hannah Harrison[d], Sabrina H. Rossi[e], Stephen J. Sharp[f], Carol Coupland[g], Julia Hippisley-Cox[h], Simon J. Griffin[i], Tobias Klatte[j], Grant D Stewart[k]

[a]The Primary Care Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK, jau20@medschl.cam.ac.uk

[b]School of Clinical Medicine, University of Cambridge, Cambridge UK, ll527@cam.ac.uk

[c]School of Clinical Medicine, University of Cambridge, Cambridge UK, ler39@cam.ac.uk

[d]The Primary Care Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK, hh504@medschl.cam.ac.uk

[e]Department of Oncology, University of Cambridge, Addenbrooke's Hospital, Cambridge Biomedical Campus, Hills Road, Cambridge, UK, sr725@cam.ac.uk

[f]MRC Epidemiology Unit, University of Cambridge, Cambridge, UK, sjs207@cam.ac.uk

[g]School of Medicine, University of Nottingham, Nottingham, UK, Carol.Coupland@nottingham.ac.uk

[h]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK, julia.hippisley-cox@phc.ox.ac.uk

[i]The Primary Care Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK, profgp@medschl.cam.ac.uk

[j]Royal Bournemouth Hospital, Bournemouth, UK, tobias.klatte@gmx.de

[k]Department of Surgery, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK, gds35@cam.ac.uk

* Equal contributors

Corresponding author

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. Tel: 01223 748693. Email: jau20@medschl.cam.ac.uk

Abstract

Objective: To systematically identify and compare the performance of prognostic models providing estimates of survival or recurrence of localised 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 01/01/2000-12/12/2019 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 had undergone surgical resection for localised RCC. For each outcome we summarised the discrimination of the each model using the C-statistic and performed multivariate random-effects meta-analysis of the logit transformed C-statistic to rank the models.

Results

From 13,549 articles, 57 included data on the performance of 22 models in external populations. C-statistics ranged from 0.59-0.90. Several risk models have been assessed in two or more external populations and have similarly high discriminative performance. For RFS, these are the Sorbellini, Karakiewicz, Leibovich and Kattan models, with UISS also in European/US populations. All have C-statistics ≥0.75 in at least half of the validations. For CSS, they are the Zisman, SSIGN, Karakiewicz, Leibovich and Sorbellini models (C-statistic ≥0.80 in at least half of the validations), and for OS they are the Leibovich, Karakiewicz, Sorbellini and SSIGN models. For all outcomes the models based on clinical features at

presentation alone (Cindolo and Yaycioglu) have consistently lower discrimination. Estimates of model calibration were only infrequently included but most underestimated survival.

Conclusion

Several models have 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

Introduction

International guidelines recommend that the surveillance of individuals with localised clear cell renal cell cancer (ccRCC) should be stratified according to the risk of developing recurrence. The American Urological Association (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 ESMO giving examples of UISS and SSIGN and the EAU citing the UISS, Leibovich, and GRANT models as being the current most relevant prognostic models for ccRCC. The decision of 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 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 following ccRCC treated with surgery with curative intent.

Materials and methods

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

Search strategy

We systematically searched Medline, EMBASE and the Cochrane Library from 01/01/2000-12/12/2019 using a combination of subject headings incorporating 'kidney cancer/renal cell cancer', 'recurrence/survival/prognosis' and 'prediction/model/score' (Supplementary Tables 1-2). 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 report a quantitative measure of the performance of one or more risk model(s) that include(s) 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 localised RCC. The outcomes of interest were drawn from the DATECAN guidelines for time-to-event end points 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 over-estimates of performance due to over-fitting, 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 localised 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% were additionally checked by a third. For studies that reported the step-wise 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 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 (Supplementary Methods 1)[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, CCS and OS). The discrimination for each model was summarised 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 (REML) 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 (BoS), 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 0-1 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 between 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 summarised descriptively.

Results

Our search identified 13,549 articles. 75 met our inclusion criteria (Supplementary Figure 1 and Supplementary Table 3). 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. 57 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 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 Supplementary Tables 4-6. Of the 150 validations assessed (69 RFS, 38 CSS and 43 OS), 95 were rated high RoB, 49 unclear RoB (typically due to a lack of clear reporting) and 6 low RoB. Across the four domains assessed (Supplementary Methods 1), 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[21,22,31–40,23,41,24–30]. The majority include pathological or clinical prognostic factors that are likely to be available in routine clinical practice. Two include genetic risk factors (Wei 2019[26] and the Recurrence score[24]), one includes molecular markers (Klatte 2009[23]) and five include biochemical markers (e.g. albumin and C-reactive protein (CRP)) that may be available in some settings (CONUT[36], GPS[39], mGPS[30], PNI[41] and Chen 2017[21]).

Recurrence-free survival

Discrimination

36 studies reported 69 C-statistics for external validations of 19 models for recurrence free survival from surgery (Supplementary Table 4). The median duration of follow-up was reported for 59 of the external validations and ranged from 33-128 months, with most (n=41/59) having a median follow-up period between 60-90 months. The discriminative performance within all the external validations for the 19 models is shown in Figure 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 is substantial variation in discriminative performance both between different models and between different studies assessing the same model (Figure 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 clear cell RCC

were able to explain that heterogeneity (Supplementary Table 5). The high heterogeneity also persisted in sub-group 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 (Figure 4a). Those eight models all had a SUCRA≥0.6 (Table 2 and Supplementary Figure 2).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) include pathological or symptom prognostic markers that are likely to be routinely available and have been validated in at least two external populations. Jeong 2017 is the only model to also include age. The other three include either genetic markers (the Recurrence score), single nucleotide polymorphisms (SNPs) (Wei 2019) or molecular markers (Klatte 2009) not currently available in clinical practice. These three, as well as the model by Jeong 2017, have 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 have comparatively poor discrimination (SUCRA 0.1 and 0.3 and reported C-statistics between 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, is 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 sub-groups are presented in Supplementary Tables 8 and 9, respectively. Except for the UISS score that performed better in European/US populations, the results were similar to 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 post-surgery. There was no significant difference in 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 the maximum departure of predicted from observed outcomes 4%, 17%, 11% and 15% respectively. Beisland *et al.* 2015[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 2011-2014, Vasudev *et al.* 2019[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

15 studies (Supplementary Table 6) 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 60-90 months. As for RFS there is substantial variation in the C-statistics (Figure 2). Seven risk models, however, appear to perform better than others, with a C-statistic ≥0.80 in at least half of the studies in which they have 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, Supplementary Figure 3) incorporating direct comparisons (Figure 4b).

Of these, the four models with the highest SUCRA (≥0.7) are the only three models developed primarily for estimating CSS (Zisman, SSIGN and Karakiewicz) and the Klatte 2009 model that includes molecular markers but that has only been externally validated in one cohort. The Leibovich and Sorbellini models, originally developed for RFS, are also in this group, along with the mGPS model which was originally developed for colorectal cancer prognosis and includes CRP and albumin but has 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, have 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, is 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 (Supplementary Table 8).

In addition to the 5-year CSS from the time of surgery, Fu et al, 2015[56] report 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 1-5 years since 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 2-5 years post- surgery.

Calibration

Only the study by Tan 2011 assessed calibration. As for RFS in the same study, all four models assessed (Karakiewicz, Leibovich, Kattan and Sorbellini) had reasonable calibration graphically[47].

Estimates of survival for risk groups

Seven studies[19,49,50,57–60] reported the probability of CSS between 1-10 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 the observed probability of survival fell moving from the low-risk to high-risk groups.

Overall survival

Discrimination

20 studies (Supplementary Table 7) reported the discrimination of 43 external validations of models for overall survival. As for RFS and CSS, heterogeneity was high (I2 up to 87.5%) so estimates were not pooled. The reported C-statistics for different models ranged from 0.59 to 0.90 (Figure 3).The model with the highest discrimination in any validation (C-statistic 0.90 (0.80-0.95)) is the model developed by Chen which includes pathological T-stage along with three biochemical ratios. That model, however, has only been validated in one small population (23 cases of RCC) from the same hospital as the model was developed in. 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, have the lowest discrimination (C-statistics 0.62-0.70 and 0.59-0.62, respectively). Despite being developed for OS, the UISS model also has 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 is reflected in the multivariate analysis where the UISS model has by a SUCRA of 0.3 and, together with the Chen model, the four highest ranking models with SUCRA≥6 are the Leibovich, SSIGN, Karakiewicz and Sorbellini models. There is little to distinguish between those four, with all 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 (Supplementary Table 9).

Calibration

Two studies reported data on calibration. As for RFS and CSS, the study by Tan 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.* 2008[62] found that the UISS model significantly (Likelihood Ratio Test p<0.0001) underestimated OS, particularly at the extremes. The difference was mainly due 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 1-5 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

40 studies externally validated pre-existing risk models and also investigated the improvement in the performance of these models when additional prognostic markers were incorporated (Supplementary Table 10). 35 evaluated additional prognostic markers for RFS, 3 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 hybridisation, 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 perform 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 the ECOG 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 are also several models that have similar performance but have only been assessed in one external population so further validation studies are required. These include models which use genetic risk markers (Recurrence score, Wei 2009), molecular markers (Klatte), biochemical markers (Chen 2017 and mGPS) and age (Jeong 2017). 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 performs 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.

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 individualised 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 is 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 between 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 standardisation over data collection, potential differences in reporting and collection methods both between centres and over time, and a lack of centralised pathological review. The recruitment periods of many of the studies also began over twenty 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 robotic partial nephrectomy, however, is unlikely to have significantly impacted on survival estimates as current data suggest that there are no differences in oncological outcomes following open partial nephrectomy, laparoscopy partial nephrectomy or robotic partial nephrectomy[66–68]. Further validation in contemporary cohorts, ideally from large prospective studies, are however 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 localised and locally advanced RCC[71]. While 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 localised 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 individualised 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.

Funding

JUS was funded by a Cancer Research UK Prevention Fellowship (C55650/A21464). The University of Cambridge has received salary support in respect of SJG from the NHS in the East of England through the Clinical Academic Reserve. SHR is funded by a Cancer Research UK Clinical PhD Fellowship. GDS is supported by the Renal Cancer Research Fund, The Mark Foundation for Cancer Research, the Cancer Research UK Cambridge Centre [C9685/A25177] and 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. HH was supported by a National Institute of Health Research Methods Fellowship (RM-SR-2017-09-009) and is now supported by a National Institute of Health Research Development and Skills Enhancement Award (NIHR301182). SJS 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; and 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 National Institute for Health Research or the Department of Health and Social Care.

Financial disclosure

GDS 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 Access and Responsibility

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.

Ethical approval

Not required.

Data sharing

All data used in this study are publically available in the primary articles.

References

1. Donat SM, Diaz M, Bishoff JT, Cole- JA, Dahm P, Derweesh IH, et al. American Urological Association (AUA) Guideline: Follow-up for clinically localized renal neoplasms. AUA Clin Guidel. 2013;1–33.

2. National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in Oncology: Kidney Cancer. 2021.

3. Escudier B, Porta C, Schmidinger M, Rioux-Leclercq N, Bex A, Khoo V, 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, Bex A, Capitanio U, Gilles R., et al. EAU guidelines on renal cell carcinoma [Internet]. 2021 [cited 2021 Aug 20]. Available from: https://uroweb.org/guideline/renal-cell-carcinoma/

5. Sun M, Shariat SF, Cheng C, Ficarra V, Murai M, Oudard S, 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. Springer Berlin Heidelberg; 2018;36:1943–52.

7. Choueiri T., Klaassen Z. 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;

8. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;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, Joniau S, Mulders P, Powles T, 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, Whiting PF, Westwood M, Collins GS, 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. DPC Nederland; 2011;11:255–70.

15. White IR. Multivariate random-effects meta-analysis. Stata J. DPC Nederland; 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. Oxford Academic; 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. SAGE Publications Ltd; 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. Pergamon; 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. Nature Publishing Group; 2012;106:279–83.

20. Vasudev NS, Hutchinson M, Trainor S, Ferguson R, Bhattarai S, Adeyoju A, et al. UK Multicenter

Prospective Evaluation of the Leibovich Score in Localized Renal Cell Carcinoma: Performance has Altered Over Time. Urology. Elsevier Inc.; 2020;136:162–8.

21. Chen Z, Shao Y, Yao H, Zhuang Q, Wang K, Xing Z, 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 JM, Shin SJ, Lee JB, Song C, Go H, 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, Shuch B, Said JW, Riggs SB, 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, Maddala T, Zhou M, Aydin H, 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, Reuter V, Motzer R, Goetzl M, 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, Zhao HW, Chen ZH, Liao B, 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. Elsevier Ltd; 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.

29. Zisman A, Pantuck AJ, Wieder J, Chao DH, Dorey F, Said JW, 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. Int J Colorectal Dis; 2007;22:881–6.

31. Cindolo L, De La Taille A, Messina G, Romis L, Abbou CC, Altieri V, et al. A preoperative clinical prognostic model for non-metastatic renal cell carcinoma. BJU Int. BJU Int; 2003. p. 901–5.

32. Zisman A, Pantuck AJ, Dorey F, Said JW, Shvarts O, Quintana D, 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, DiPaola RS, Manola J, Taguchi S, 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. Oxford University Press; 2017;28:2747–53.

34. Karakiewicz PI, Briganti A, Chun FKH, Trinh QD, Perrotte P, Ficarra V, et al. Multi-institutional validation of a new renal cancer-specific survival nomogram. J Clin Oncol. J Clin Oncol; 2007;25:1316–22.

35. Dall'Oglio MF, Ribeiro-Filho LA, Antunes AA, Crippa A, Nesrallah L, Gonçalves PD, 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, G P, Gonzalez P, Gonzalez B, Mancha A, 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, George DJ, Pantuck AJ, Patel A, et al. Adjuvant Sunitinib in High-Risk Renal-Cell Carcinoma after Nephrectomy. N Engl J Med. Massachusetts Medical Society; 2016;375:2246–54.

38. Leibovich BC, Blute ML, Cheville JC, Lohse CM, Frank I, Kwon ED, et al. Prediction of progression

after radical nephrectomy for patients with clear cell renal cell carcinoma: A stratification tool for prospective clinical trials. Cancer. 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. Nature Publishing Group; 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, Vertosick EA, Cianflone F, Carenzi C, 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, Komaru A, Sazuka T, Kawamura K, 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, Benfante NE, Kou L, Yu C, 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 JF, Chretien Y, Dufour B, Thiounn N, 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 MH, Li H, Choong CV, Chia KS, Toh CK, Tang T, 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, Pflanz S, Roigas J, Klotz T, 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, Knoll N, Roigas J, Hoschke B, 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, Zhang Z, Ma B, Jin J, 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, Liu Y, Lin Z, Wang G, 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.e9-68.e16.

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, Fu Q, Chen L, Lin Z, 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, Jesche J, Groselj-Strele A, Kampel-Kettner K, et al. Prognostic value of the leibovich prognosis score supplemented by vascular invasion for clear cell renal cell carcinoma. J Urol. Elsevier Inc.; 2012;187:834–9.

55. Seles M, Posch F, Pichler GP, Gary T, Pummer K, Zigeuner R, 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. American Urological

Association Education and Research, Inc.; 2017;198:1247–52.

56. Fu Q, Chang Y, An H, Fu H, Zhu Y, Xu L, 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 KR, Bleumer I, Pantuck AJ, Kim HL, Dorey FJ, Janzen NK, 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, Yoshikawa Y, Takai T, Uchimoto 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. Springer International Publishing; 2017;24:2787–93.

59. Morgan TM, Mehra R, Tiemeny P, Wolf JS, Wu S, Sangale Z, 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. European Association of Urology; 2018;73:763–9.

60. Ficarra V, Novara G, Galfano A, Brunelli M, Cavalleri S, Martignoni G, 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. Stat Med; 2000. p. 3401–15.

62. Cindolo L, Chiodini P, Gallo C, Ficarra V, Schips L, Tostain J, 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, Capitanio U, Tian Z, Cortellini A, 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, Gamble C, Dodd S, Smyth R, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ. British Medical Journal Publishing Group; 2010;340:637–40.

65. Dwan K, Altman DG, Clarke M, Gamble C, Higgins JPT, Sterne JAC, 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. Public Library of Science; 2014;11:e1001666.

66. Chang KD, Abdel Raheem A, Kim KH, Oh CK, Park SY, Kim YS, 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, Vaessen C, Grassano Y, Benoit T, 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. European Association of Urology; 2015;67:891–901.

69. Thompson RH, Leibovich BC, Lohse CM, Cheville JC, Zincke H, Blute ML, 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, Zaid HB, Boorjian SA, Frank I, et al. Predicting Oncologic Outcomes in Renal Cell Carcinoma After Surgery. Eur Urol. European Association of Urology; 2018;73:772–80.

71. Correa AF, Jegede OA, Haas NB, Flaherty KT, Pins MR, Adeniran A, 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. European Association of Urology;

2021;80:20–31.

72. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, 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, Flaherty KT, Pins MR, Messing EM, et al. Predicting renal cancer recurrence: Defining limitations of existing prognostic models with prospective trial-based validation. J Clin Oncol. 2019;37:2062–71.

Figure legends

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

Figure 2. Forest plot showing the C-statistics from individual studies for cancer specific survival (CSS).

Figure 3. Forest plot showing the C-statistics from individual studies for overall survival (OS).

Figure 4. 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 (OC). 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.

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 2017[21]	China	ccRCC	OS	1. T stage 2. Neutrophil to lymphocyte ratio (NLR) 3. Monocyte to lymphocyte ratio (MLR) 4. Albumin to globulin ratio (AGR)	Nomogram giving continuous quantification of risk	Potentially
Cindolo[31]	France, Italy	RCC	RFS	1. Clinical size 2. Clinical presentation (symptomatic vs asymptomatic)	Recurrence risk formula (RRF) = (1.28 x presentation (asymptomatic = 0; symptomatic = 1) + (0.13 x 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 NSCLC	OS	1. CRP 2. Albumin	**Score** Elevated CRP (>10 mg/L) and hypoalbuminemia (<35 g/L) = 2 Elevated CRP (>10 mg/L) or hypoalbuminemia (<35 g/L) = 1 CRP ≤10mg/L and albumin ≥35g/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 2017[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[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[27]	USA	RCC	RFS	1. Pathological tumour stage 2. Tumour size 3. Histology 4. Symptoms	Nomogram giving continuous quantification of risk	Y
Klatte 2009[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[38]	USA	ccRCC	RFS	1. Pathologic T stage 2. Regional lymph node status (N stage) 3. Tumour size 4. Nuclear grade 5. Histologic tumour necrosis	Score range 0-11 **Low risk:** score 0-2 **Intermediate risk:** score 3-5 **High risk:** score ≥ 6	Y
mGPS[30]	UK	Colorectal cancer		1. CRP 2. Albumin	**Score** Elevated CRP (>10 mg/L) and hypoalbuminemia (<35 g/L) = 2 Elevated CRP (>10 mg/L) and albumin ≥35g/L = 1 CRP ≤10mg/L = 0	Potentially
PNI[41]	Japan	GI cancer	Postop complications	1. Serum albumin level 2. Total lymphocytes count	PNI = (10 x serum albumin level (g/100 ml)) + (0.005 x total lymphocyte count/mm3 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. Microvascular invasion (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[25]	USA	ccRCC	RFS	1. 2002 TNM stage 2. Tumour size (cm) 3. Fuhrman grade 4. Necrosis	Nomogram giving continuous quantification of risk	Y

				5. Vascular invasion 6. Clinical presentation a. Incidental asymptomatic, b. Locally symptomatic c. Systemically symptomatic		
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 cancer specific survival	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
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, FG 1-2, PS 0 **Group II**: Any other TNM stage 1; TNM stage 2; TNM stage 3, any FG, PS 0; TNM stage 3, FG 1, PS $\geqslant$1 **Group III:** TNM stage 3, FG 2-4, PS $\geqslant$1; TNM stage 4, FG 1-2, PS 0 **Group IV:** TNM stage 4, FG 3-4, PS 0; TNM stage 4, FG 1-3, PS $\geqslant$1 **Group V:** TNM stage 4, FG 4, PS $\geqslant$1	Y
Wei 2019[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[40]	USA	RCC	RFS	1. Pre-operative clinical tumour size 2. Presentation	Recurrence risk (R_{rec}) = 1.55 x presentation (0-1) + 0.19 x clinical size (in cm).	Y

				a. symptomatic b. asymptomatic	**Low risk:** R_{rec} score ≤ 3.0 **High risk:** R_{rec} score >3.0	
Zisman[29]	USA	RCC	CCS	1. 1997 T classification 2. Fuhrman grade 3. ECOG PS	**Low risk:** pT1N0M0, FG 1-2, PS 0; **Intermediate risk:** Any other N0M0 **High risk:** T3N0M0, FG >1, PS ⩾1; Any pT4N0M0	Y

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	---	---	---

*****H – High risk of bias, U – Unclear risk of bias, L – Low risk of bias

******Excluded from multivariate analysis as only assessed in one population with no other risk models

*** Events not reported for two studies

Table 3.

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

* Although median follow-up only 6.1 years

Figure 1

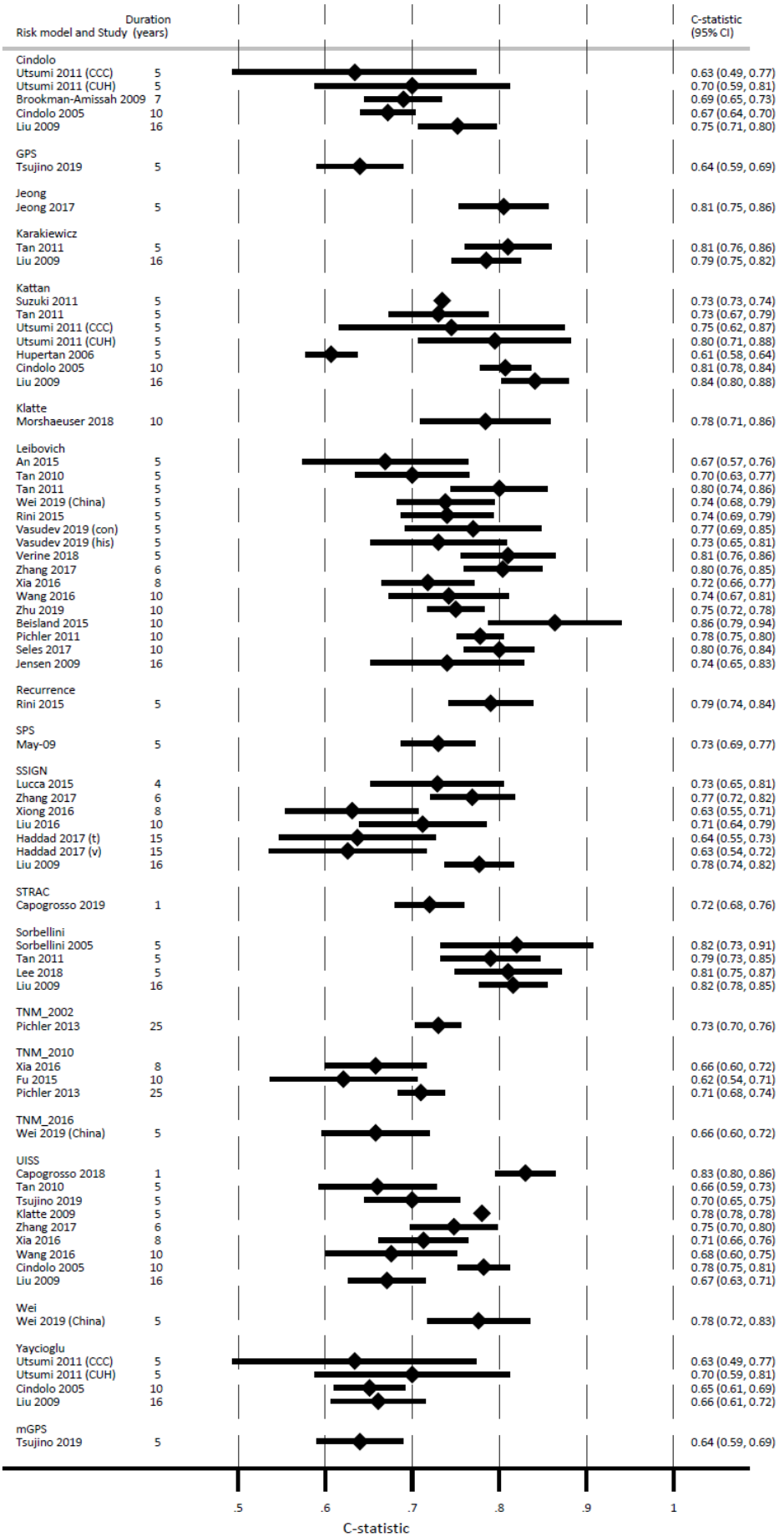

Risk model and Study
Duration (years)
C-statistic (95% CI)
Cindolo
Utsumi 2011 (CCC) 5 0.63 (0.49, 0.77)
Utsumi 2011 (CUH) 5 0.70 (0.59, 0.81)
Brookman-Amissah 2009 7 0.69 (0.65, 0.73)
Cindolo 2005 10 0.67 (0.64, 0.70)
Liu 2009 16 0.75 (0.71, 0.80)
GPS
Tsujino 2019 5 0.64 (0.59, 0.69)
Jeong
Jeong 2017 5 0.81 (0.75, 0.86)
Karakiewicz
Tan 2011 5 0.81 (0.76, 0.86)
Liu 2009 16 0.79 (0.75, 0.82)
Kattan
Suzuki 2011 5 0.73 (0.73, 0.74)
Tan 2011 5 0.73 (0.67, 0.79)
Utsumi 2011 (CCC) 5 0.75 (0.62, 0.87)
Utsumi 2011 (CUH) 5 0.80 (0.71, 0.88)
Hupertan 2006 5 0.61 (0.58, 0.64)
Cindolo 2005 10 0.81 (0.78, 0.84)
Liu 2009 16 0.84 (0.80, 0.88)
Klatte
Morshaeuser 2018 10 0.78 (0.71, 0.86)
Leibovich
An 2015 5 0.67 (0.57, 0.76)
Tan 2010 5 0.70 (0.63, 0.77)
Tan 2011 5 0.80 (0.74, 0.86)
Wei 2019 (China) 5 0.74 (0.68, 0.79)
Rini 2015 5 0.74 (0.69, 0.79)
Vasudev 2019 (con) 5 0.77 (0.69, 0.85)
Vasudev 2019 (his) 5 0.73 (0.65, 0.81)
Verine 2018 5 0.81 (0.76, 0.86)
Zhang 2017 6 0.80 (0.76, 0.85)
Xia 2016 8 0.72 (0.66, 0.77)
Wang 2016 10 0.74 (0.67, 0.81)
Zhu 2019 10 0.75 (0.72, 0.78)
Beisland 2015 10 0.86 (0.79, 0.94)
Pichler 2011 10 0.78 (0.75, 0.80)
Seles 2017 10 0.80 (0.76, 0.84)
Jensen 2009 16 0.74 (0.65, 0.83)
Recurrence
Rini 2015 5 0.79 (0.74, 0.84)
SPS
May-09 5 0.73 (0.69, 0.77)
SSIGN
Lucca 2015 4 0.73 (0.65, 0.81)
Zhang 2017 6 0.77 (0.72, 0.82)
Xiong 2016 8 0.63 (0.55, 0.71)
Liu 2016 10 0.71 (0.64, 0.79)
Haddad 2017 (t) 15 0.64 (0.55, 0.73)
Haddad 2017 (v) 15 0.63 (0.54, 0.72)
Liu 2009 16 0.78 (0.74, 0.82)
STRAC
Capogrosso 2019 1 0.72 (0.68, 0.76)
Sorbellini
Sorbellini 2005 5 0.82 (0.73, 0.91)
Tan 2011 5 0.79 (0.73, 0.85)
Lee 2018 5 0.81 (0.75, 0.87)
Liu 2009 16 0.82 (0.78, 0.85)
TNM_2002
Pichler 2013 25 0.73 (0.70, 0.76)
TNM_2010
Xia 2016 8 0.66 (0.60, 0.72)
Fu 2015 10 0.62 (0.54, 0.71)
Pichler 2013 25 0.71 (0.68, 0.74)
TNM_2016
Wei 2019 (China) 5 0.66 (0.60, 0.72)
UISS
Capogrosso 2018 1 0.83 (0.80, 0.86)
Tan 2010 5 0.66 (0.59, 0.73)
Tsujino 2019 5 0.70 (0.65, 0.75)
Klatte 2009 5 0.78 (0.78, 0.78)
Zhang 2017 6 0.75 (0.70, 0.80)
Xia 2016 8 0.71 (0.66, 0.76)
Wang 2016 10 0.68 (0.60, 0.75)
Cindolo 2005 10 0.78 (0.75, 0.81)
Liu 2009 16 0.67 (0.63, 0.71)
Wei
Wei 2019 (China) 5 0.78 (0.72, 0.83)
Yaycioglu
Utsumi 2011 (CCC) 5 0.63 (0.49, 0.77)
Utsumi 2011 (CUH) 5 0.70 (0.59, 0.81)
Cindolo 2005 10 0.65 (0.61, 0.69)
Liu 2009 16 0.66 (0.61, 0.72)
mGPS
Tsujino 2019 5 0.64 (0.59, 0.69)
.5
.6
.7
.8
.9
1
C-statistic

Figure 2

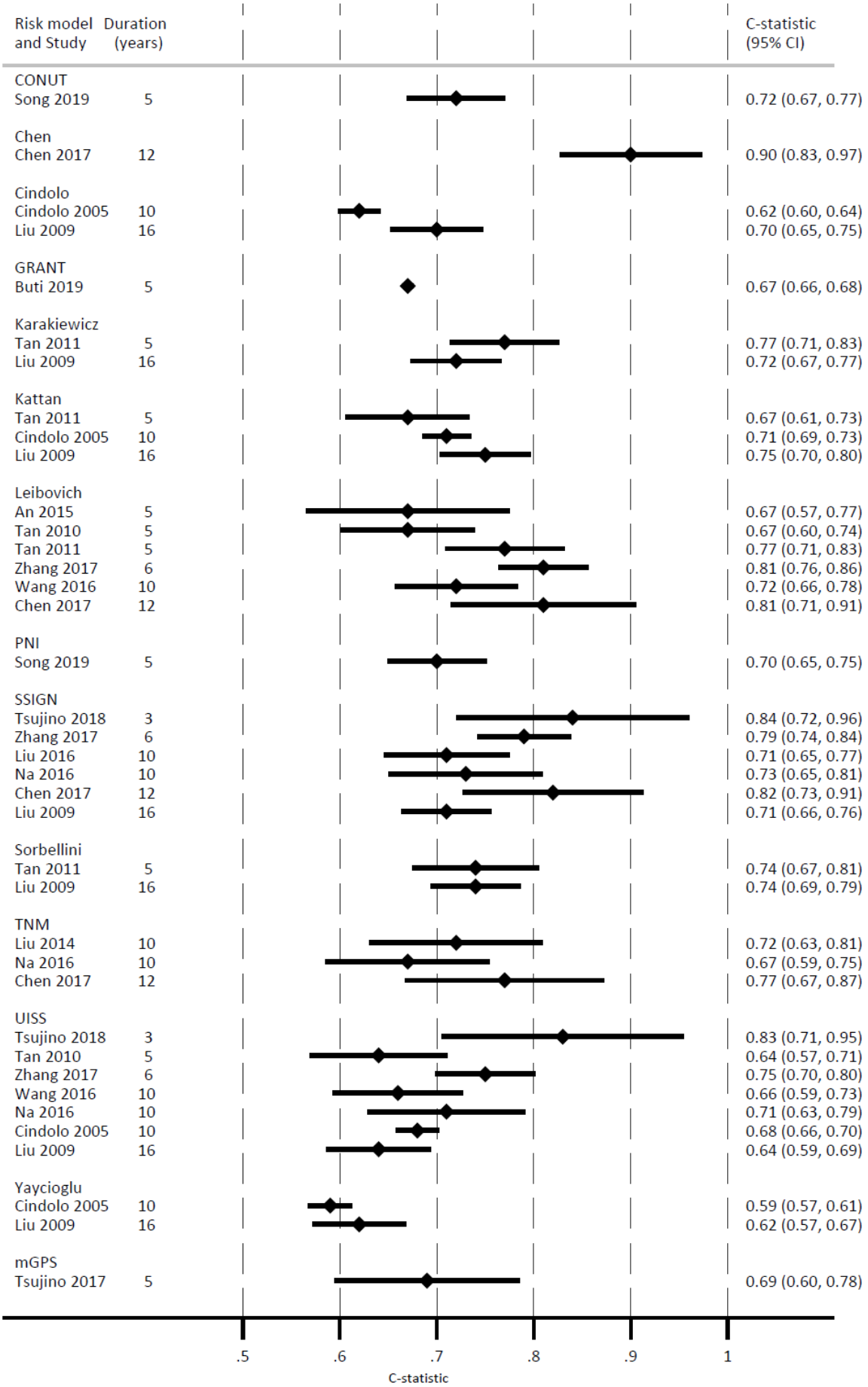

Risk model and Study
Duration (years)
C-statistic (95% CI)
CONUT
Song 2019 5 0.72 (0.67, 0.77)
Chen
Chen 2017 12 0.90 (0.83, 0.97)
Cindolo
Cindolo 2005 10 0.62 (0.60, 0.64)
Liu 2009 16 0.70 (0.65, 0.75)
GRANT
Buti 2019 5 0.67 (0.66, 0.68)
Karakiewicz
Tan 2011 5 0.77 (0.71, 0.83)
Liu 2009 16 0.72 (0.67, 0.77)
Kattan
Tan 2011 5 0.67 (0.61, 0.73)
Cindolo 2005 10 0.71 (0.69, 0.73)
Liu 2009 16 0.75 (0.70, 0.80)
Leibovich
An 2015 5 0.67 (0.57, 0.77)
Tan 2010 5 0.67 (0.60, 0.74)
Tan 2011 5 0.77 (0.71, 0.83)
Zhang 2017 6 0.81 (0.76, 0.86)
Wang 2016 10 0.72 (0.66, 0.78)
Chen 2017 12 0.81 (0.71, 0.91)
PNI
Song 2019 5 0.70 (0.65, 0.75)
SSIGN
Tsujino 2018 3 0.84 (0.72, 0.96)
Zhang 2017 6 0.79 (0.74, 0.84)
Liu 2016 10 0.71 (0.65, 0.77)
Na 2016 10 0.73 (0.65, 0.81)
Chen 2017 12 0.82 (0.73, 0.91)
Liu 2009 16 0.71 (0.66, 0.76)
Sorbellini
Tan 2011 5 0.74 (0.67, 0.81)
Liu 2009 16 0.74 (0.69, 0.79)
TNM
Liu 2014 10 0.72 (0.63, 0.81)
Na 2016 10 0.67 (0.59, 0.75)
Chen 2017 12 0.77 (0.67, 0.87)
UISS
Tsujino 2018 3 0.83 (0.71, 0.95)
Tan 2010 5 0.64 (0.57, 0.71)
Zhang 2017 6 0.75 (0.70, 0.80)
Wang 2016 10 0.66 (0.59, 0.73)
Na 2016 10 0.71 (0.63, 0.79)
Cindolo 2005 10 0.68 (0.66, 0.70)
Liu 2009 16 0.64 (0.59, 0.69)
Yaycioglu
Cindolo 2005 10 0.59 (0.57, 0.61)
Liu 2009 16 0.62 (0.57, 0.67)
mGPS
Tsujino 2017 5 0.69 (0.60, 0.78)
.5
.6
.7
.8
.9
1
C-statistic

Figure 3

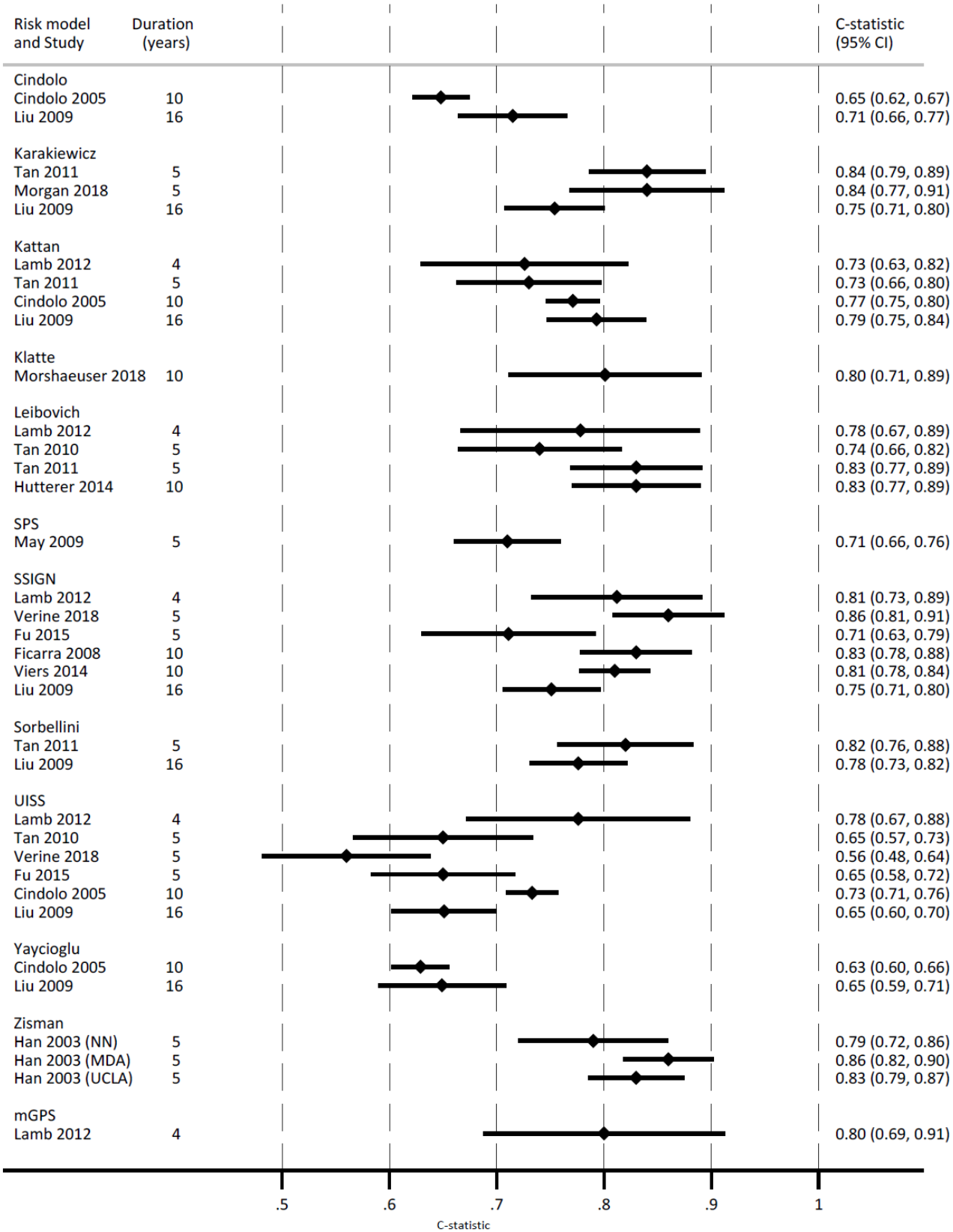
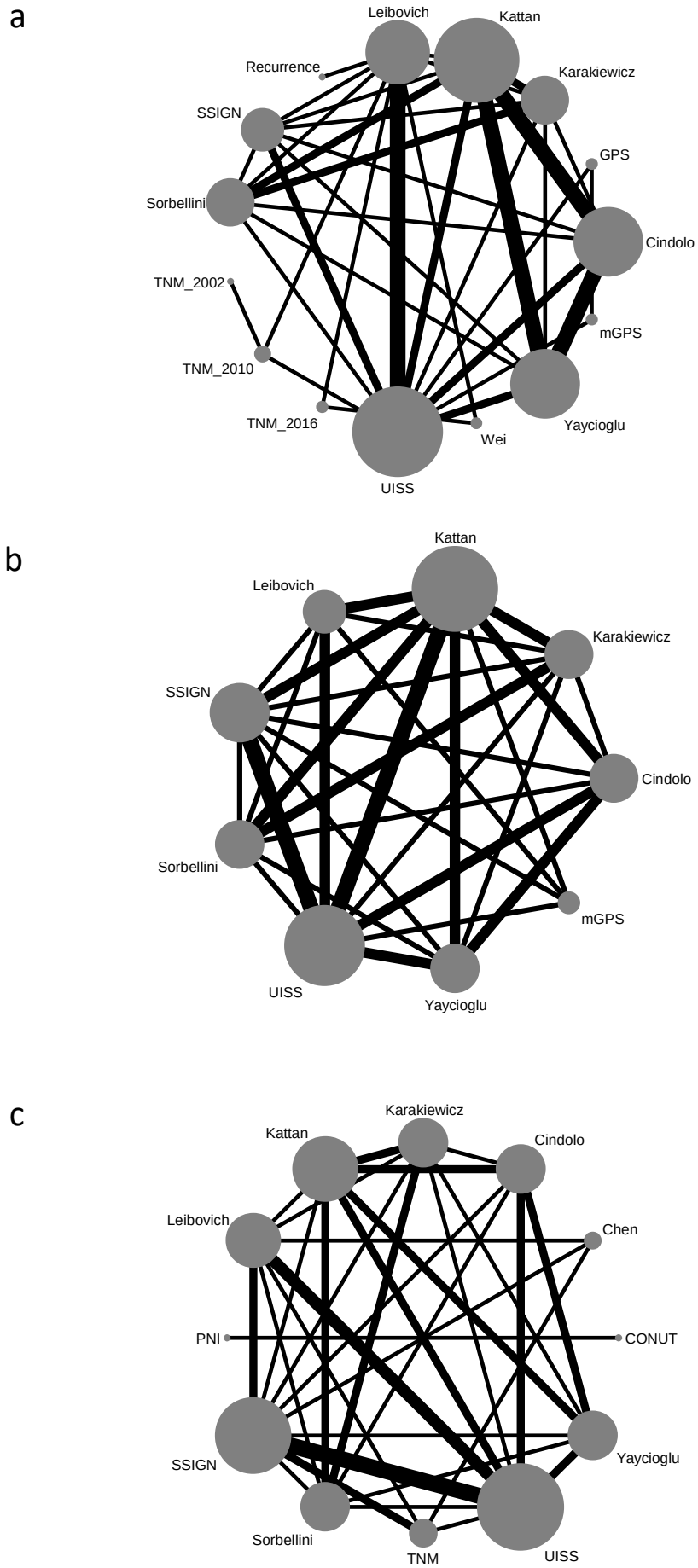

Risk model and Study
Duration (years)
C-statistic (95% CI)
Cindolo
Cindolo 2005 10 0.65 (0.62, 0.67)
Liu 2009 16 0.71 (0.66, 0.77)
Karakiewicz
Tan 2011 5 0.84 (0.79, 0.89)
Morgan 2018 5 0.84 (0.77, 0.91)
Liu 2009 16 0.75 (0.71, 0.80)
Kattan
Lamb 2012 4 0.73 (0.63, 0.82)
Tan 2011 5 0.73 (0.66, 0.80)
Cindolo 2005 10 0.77 (0.75, 0.80)
Liu 2009 16 0.79 (0.75, 0.84)
Klatte
Morshaeuser 2018 10 0.80 (0.71, 0.89)
Leibovich
Lamb 2012 4 0.78 (0.67, 0.89)
Tan 2010 5 0.74 (0.66, 0.82)
Tan 2011 5 0.83 (0.77, 0.89)
Hutterer 2014 10 0.83 (0.77, 0.89)
SPS
May 2009 5 0.71 (0.66, 0.76)
SSIGN
Lamb 2012 4 0.81 (0.73, 0.89)
Verine 2018 5 0.86 (0.81, 0.91)
Fu 2015 5 0.71 (0.63, 0.79)
Ficarra 2008 10 0.83 (0.78, 0.88)
Viers 2014 10 0.81 (0.78, 0.84)
Liu 2009 16 0.75 (0.71, 0.80)
Sorbellini
Tan 2011 5 0.82 (0.76, 0.88)
Liu 2009 16 0.78 (0.73, 0.82)
UISS
Lamb 2012 4 0.78 (0.67, 0.88)
Tan 2010 5 0.65 (0.57, 0.73)
Verine 2018 5 0.56 (0.48, 0.64)
Fu 2015 5 0.65 (0.58, 0.72)
Cindolo 2005 10 0.73 (0.71, 0.76)
Liu 2009 16 0.65 (0.60, 0.70)
Yaycioglu
Cindolo 2005 10 0.63 (0.60, 0.66)
Liu 2009 16 0.65 (0.59, 0.71)
Zisman
Han 2003 (NN) 5 0.79 (0.72, 0.86)
Han 2003 (MDA) 5 0.86 (0.82, 0.90)
Han 2003 (UCLA) 5 0.83 (0.79, 0.87)
mGPS
Lamb 2012 4 0.80 (0.69, 0.91)
.5
.6
.7
.8
.9
1
C-statistic

Figure 4


a
Leibovich
Kattan
Recurrence
Karakiewicz
SSIGN
GPS
Sorbellini
Cindolo
TNM_2002
mGPS
TNM_2010
TNM_2016
Wei
Yaycioglu
UISS
b
Kattan
Leibovich
Karakiewicz
SSIGN
Cindolo
Sorbellini
mGPS
UISS
Yaycioglu
c
Karakiewicz
Kattan
Cindolo
Leibovich
Chen
PNI
CONUT
SSIGN
Yaycioglu
Sorbellini
TNM
UISS